



Stockholm
University

Pär Nyman

Methods in Political Science
30 March 2026

Precision and statistical inference

Standard errors, conf. intervals and stat. significance

Introduction

Imagine that you have asked 500 respondents if they support the Centre Party. Can you calculate a confidence interval for the party's support in the population?

SERIOUSLY?

THIS AGAIN?

Introduction

Statistical inference is a familiar topic, but

- Commonly misunderstood.
- Regression-focus implies some news?
 - From sample to data-generating process.
 - Larger emphasis on standard errors.
 - Correlated residuals.
- It is broken.

Introduction

Rough outline

1. Why we do it
2. How we do it (mostly repetition)
 - Standard errors
 - Confidence intervals
 - T-values
 - P-values
3. Why we get it wrong

Introduction

Statistical inference for descriptive statistics

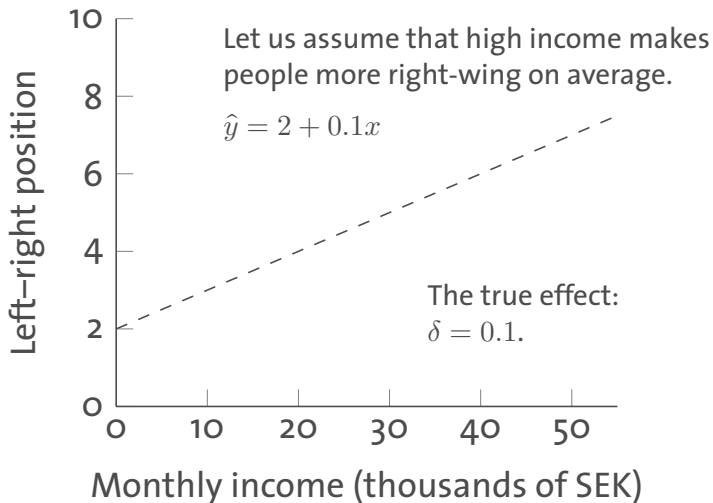
- We are interested in quantifying some characteristic for a specific population.
- We cannot study everyone so we draw a random sample.
- For each individual in this sample we can observe the characteristic of interest.
- There will be sampling errors because we did not observe the full population.
- Little need for statistical inference if you have studied the whole population.

Introduction

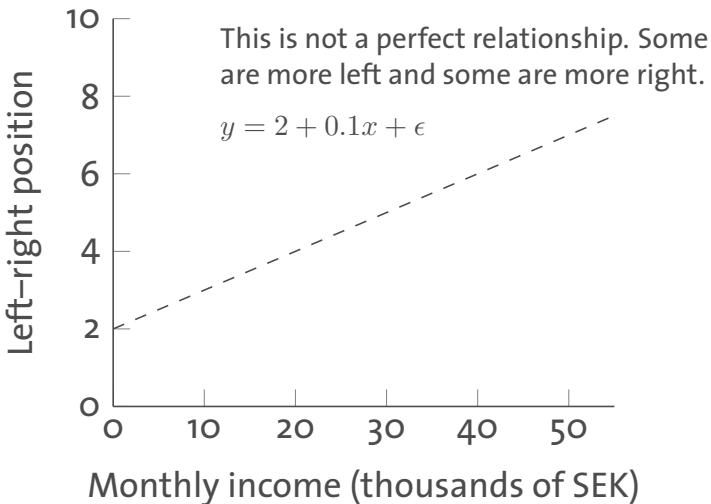
Statistical inference for causal relationships

- We still test significance and calculate confidence intervals.
- Required also for inference about the effect in the sample.
- Individual causal effects are not discernible because we cannot observe the counterfactual.
- We estimate them through comparisons within a sample.
- These estimates are afflicted with sampling errors and we therefore need statistical inference to draw conclusions about the effect also within a sample.
- We can think about it as drawing inference about the data-generating process.

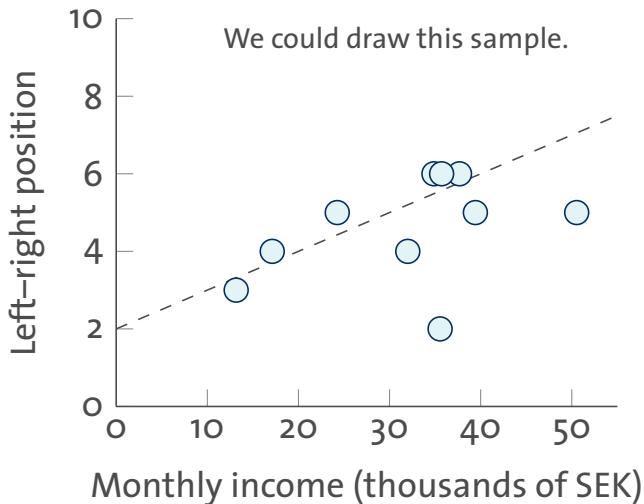
Introduction



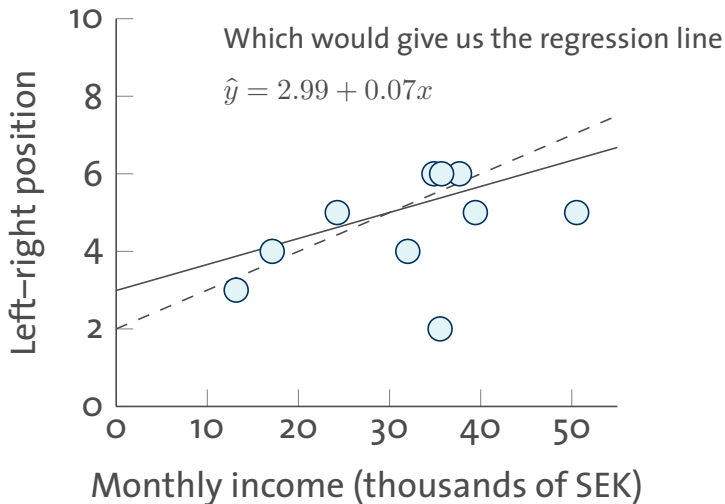
Introduction



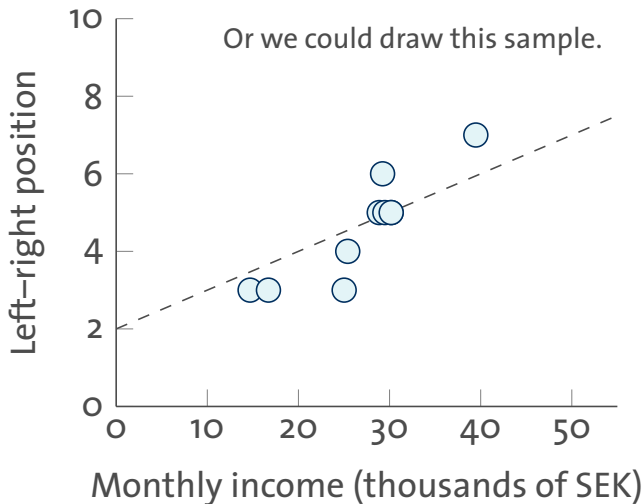
Introduction



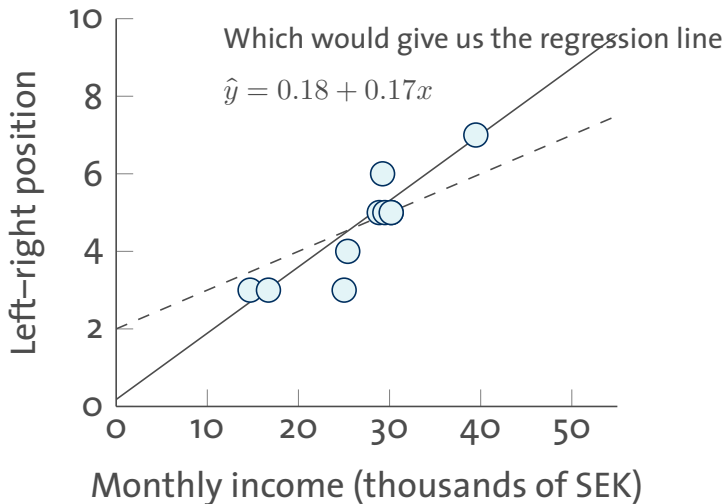
Introduction



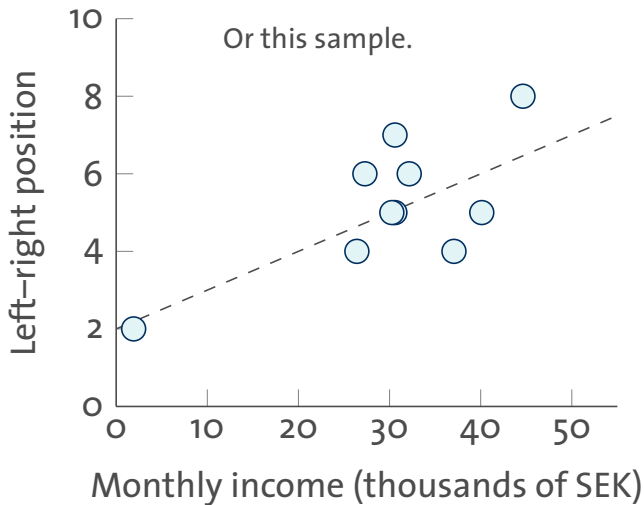
Introduction



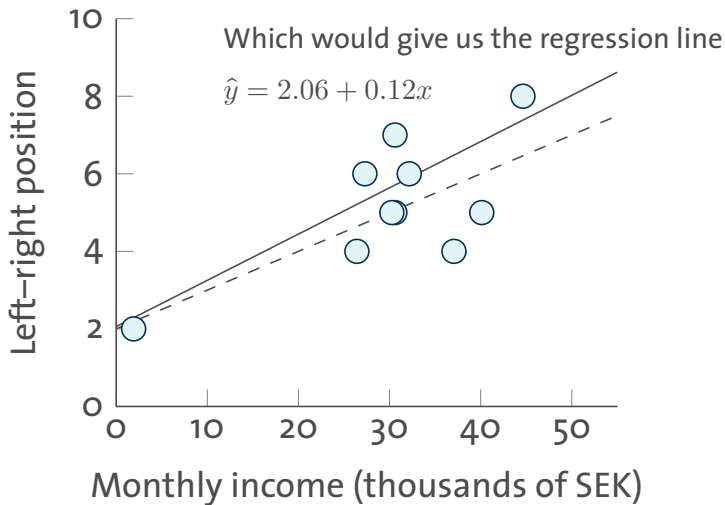
Introduction



Introduction



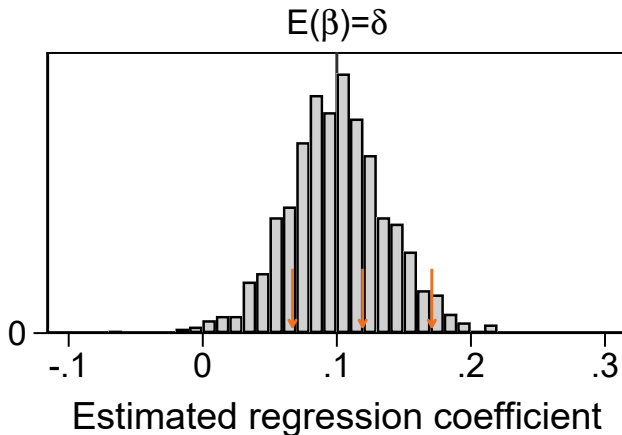
Introduction



Introduction

- In this case, the true “effect” was 0.1.
- We observe a sample and estimate it to 0.07, 0.17, 0.12, etc.
- If we kept on going, and stored the regression coefficient from each sample, we would get a sampling distribution:

Introduction



Introduction

Statistical precision

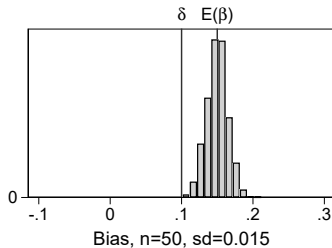
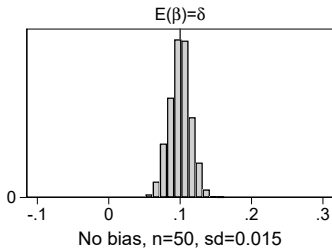
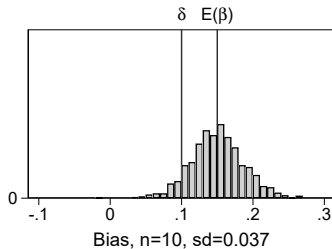
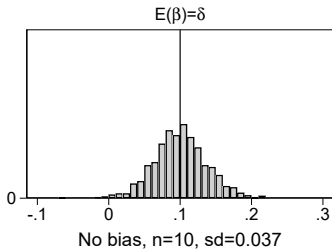
- Wide distribution = large sampling errors, bad precision.
- Narrow distribution = small sampling errors, good precision.
- We do not know where in the distribution our sample is.
- We want the sampling distribution to be narrow, and we want an accurate estimate of its dispersion.

Introduction

Statistical precision and bias

- Estimated effects can differ from the true effect because of bias and sampling errors.
- Because we made up the data-generating process, we could estimate the true effect without bias: the expected value of b , $E(\beta)$, equals the true effect δ .
- Precision = the expected distance between β and $E(\beta)$.
- Precise estimates can still be biased ($E(\beta) \neq \delta$).
- Bias is a separate issue which we do not deal with today.

Introduction



Introduction

The real world

- Both true effect and $E(\beta)$ are unknown.
 - β is our estimate.
- Sampling distribution is unknown.
 - We can assume a normal/t-distribution.
 - We can estimate its standard deviation.
- Our estimate (β) can be anywhere in the distribution.
 - We can calculate a confidence interval around our estimate that is likely to contain $E(\beta)$.

Standard errors

The standard error

- Precision is measured with a standard error.
- Approximately the expected deviation from $E(\beta)$.
- The sampling distribution's estimated standard deviation.
- The smaller, the better, as long as they are correct.

Standard errors

$$SE_{\beta} = \sqrt{\frac{1}{n-k} \frac{\sum_1^n (y_i - \hat{y})^2}{\sum_1^n (x_i - \hat{x})^2}}$$

Standard errors are affected by:

- Observations / degrees of freedom ($\frac{1}{n-k}$)
- Residual variation in y ($\sum_1^n (y_i - \hat{y})^2$)
- Unique variation in x ($\sum_1^n (x_i - \hat{x})^2$)

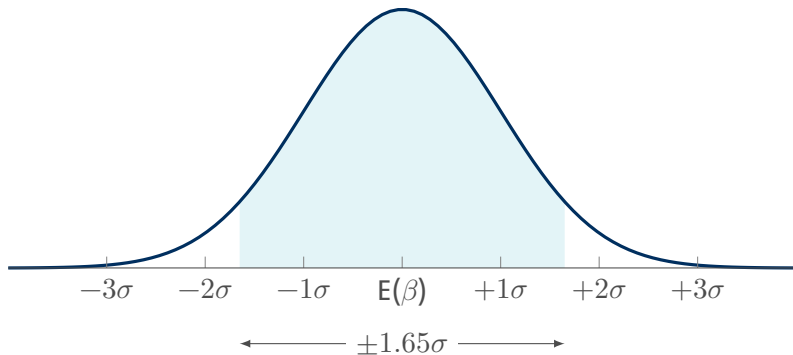
Standard errors

Three ways to improve precision / reduce the SE

- Increase the number of observations.
 - Gather more data
 - Remove/substitute variables
 - Imputation
- Decrease the residual/unexplained variation in y .
 - Add relevant independent variables
 - Get better measure of y
- Increase the unique/unexplained variation in x .
 - Remove correlated variables
 - Get better measure of x

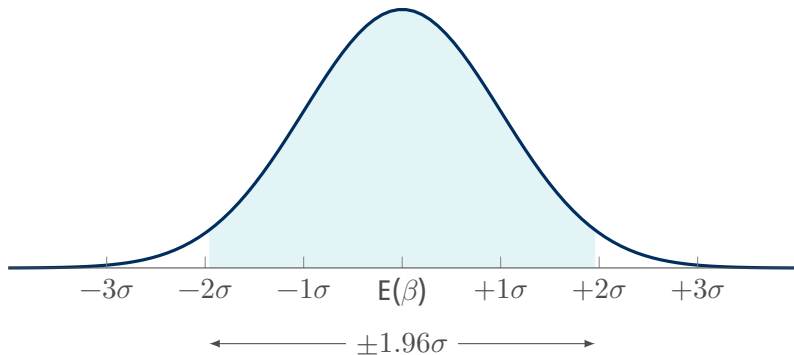
Inference

Normal distribution: 90 percent within 1.65 standard deviations from the mean.



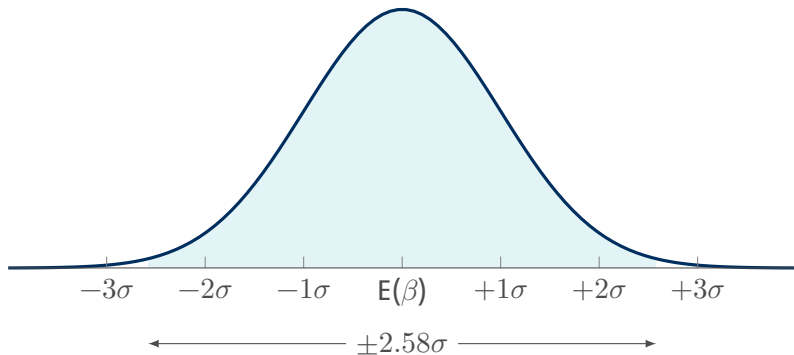
Inference

Normal distribution: 95 percent within 1.96 standard deviations from the mean.



Inference

Normal distribution: 99 percent within 2.58 standard deviations from the mean.



Inference

The confidence interval

- A confidence interval is a range of values corresponding to a certain level of confidence..
- 90/95/99 *a priori* probability obtaining b within 1.65/1.96/2.58 standard errors from $E(b)$.
- A confidence interval $\beta \pm 1.96SE_\beta$ will include $E(\beta)$ 95 percent of the time.
 - But only if the standard error is correct.
 - $E(\beta)$ only equals the true effect if unbiased / model is correct.

Inference

T-values

- By dividing the slope coefficient with its standard error we get a t-value ($t = b/se$).
- An absolute t-value larger than 1.65/1.96/2.58 means that the confidence interval does not include zero at the 90/95/99 percent confidence level (10/5/1 percent risk).
- This is synonymous with being statistically significant (assuming the standard null hypothesis).
- I never report t-values, but they are often used in teaching.

Inference

P-values

- P-values are obtained by comparing the t-value to 1.65/1.96/2.58 and other critical values. They say the risk level at which zero (H_0) would be at the end of the CI.
- Typically, it is the probability of obtaining a coefficient at least as large as β under the assumption that $E(\beta) = 0$.
- Assuming our model is correct, it is a test of whether we can rule out that the true effect is zero.

Inference

Statistical significance

- Everyone knows to examine statistical significance.
 - Does the CI include 0?
 - Is $|t|$ larger than 1.96?
 - Is $p < 0.05$?
- Commonly interpreted as a test of if there is an effect.
 - But that is assuming $E(\beta)$ equals the true effect.
 - Rejecting the null hypothesis indicates that the model or the null hypothesis is incompatible with the data.

Problems

Problems

- Our standard errors can be wrong.
- Some problems with significance-testing.
 - Misconceptions of individuals.
 - Multiple comparisons and publication bias.

Robust standard errors

- Standard errors are estimates and can be wrong.
- Conventional standard errors assume that the residuals have constant variance (homoskedasticity).
- That is, the dispersion around the regression line is not smaller for some values of x and larger for others.
- Robust standard errors relax that assumption by giving more weight to residuals from observations that matter more for estimating β – that is, observations with greater leverage, or values of x that are further from its mean.

Clustered standard errors

- Both conventional and robust standard errors assume independent residuals, but violations are common:
 - Repeated observations of the same unit over time.
 - Clustered treatments, such as reforms implemented at the school or municipality level.
 - Spatial correlation across neighbouring units.
- In such cases the number of observations exaggerates the true amount of information.
- Unless accounted for, such violations typically bias the standard errors downwards.

Clustered standard errors

- The standard solution is to adjust the standard errors for correlation within clusters.
- This corresponds to assuming that observations are independent across clusters but may be correlated within.
- The clusters are groups defined by the researcher.
 - Sometimes this is straightforward: the units in a panel or the groups in a clustered experiment.
 - Other times it is less obvious, and different clustering choices may be defensible.
- A common rule of thumb is to have at least 40–50 clusters, but this is often violated in practice.
- All common implementations of clustered standard errors also allow for heteroskedasticity within clusters.

Clustered standard errors

- We will talk more about clustering during the week on panel data.

Misconceptions about significance

Scepticism against significance testing

- You are taught to examine statistical significance.
- Meaningful analysis – if the point-estimate is not larger than what could be expected by chance, then we should not put too much weight to it.
- But this practise has caused us many problems and led many to question statistical tests at large.
 - Partly because how researchers interpret them.
 - Partly because how results are selected.

Misconceptions about significance

Keep the following in mind

- Not significant does not validate the null hypothesis.
- The difference between significant and not significant is not itself statistically significant.
- Size matters.
- Statistical significance does not validate the model.
- Opinions about the probability/likelihood/certainty of a hypothesis cannot be based on statistics alone.

It helps to remember that statistical significance is a property of the test, not of the phenomenon being studied.

Multiple comparisons

p-values for a single test

- p tells us the probability of obtaining a result at least this far from the null-hypothesis if (the model is correct and) the null-hypothesis is true.
- Rejecting H_0 when $p < 0.05$ gives us a 5 percent risk of Type I errors (false positives).
- For a single test, this is usually a satisfactory level of confidence.

Multiple comparisons

The probability of at least one Type I error increases with the number of tests.

- If we invent a new drug, and test it on 100 symptoms, we would (in expectation) find 5 statistically significant effects, even if the drug is worthless.
- If public opinion is stable, we will on average have 2.4 statistically significant changes each month according to the Swedish polling companies ($8 \times 6 \times 0.05$).
- The more specifications you run, the more false discoveries you will get. Problematic if we focus our attention to the statistically significant results.

Multiple comparisons

Solutions

- Adjust p-values for number of test.
 - Trade-off between false positives and false negatives.
 - Bonferroni correction is easy and well-known (but not the best).
- Replicate findings on independent data.
- Focus on aggregates (combine polls, look at main effects)
- Pre-registration of analyses.
- Or just be more cautious when interpreting results from an analysis that involves multiple comparisons. Treat it as exploratory?

Multiple comparisons

Publication bias

- The outcome influences if, and where, a study is published.
 - Rejections at the journals.
 - Never submitted.
 - Models adjusted until $p < .0.05$.
- Consequence: more false positives than 5 percent and upwards bias in estimated effects (Type M-errors).
- Similar arguments can be made about situations outside academia (e.g., what news that are published).
- Likely that AI-generated papers will inflate the problem.

Multiple comparisons

Solutions

- Stricter conventions for p-values.
 - Reduces the number of false positives.
 - Increases the size of Type M-errors.
- Replication studies.
- Include non-published research in meta-studies.
- Rely less on journals and more on AI to synthesize findings.
- Require pre-registration of analyses.
- More generally: always keep in mind the possibility of rare events, especially when you evaluate something that has been selected from a large number of tests.

Move away from statistical significance?

Move towards avoiding the dichotomization

- There is always uncertainty. Describe it.
- The usual null hypothesis is not always that interesting.
 - Sometimes we should make another test.
 - Confidence intervals are often better.
 - Sometimes we need to look at a broader picture than the single comparison or estimate.
- This is not an excuse for giving disproportional weight to imprecise estimates.
- Sometimes we must use our personal judgment.

Concluding

Most important take aways

- Statistical significance remains important and you must know how to check it.
 - Is the p-value below 0.05?
 - Does the confidence interval exclude zero?
 - Is the slope coefficient twice the size of its standard error?
- But do not over-emphasize it. Put more weight on the standard error or the confidence interval.
- Be aware that large sampling errors are likely if interesting or significant results are given priority.

Concluding

Thank you for listening!