# Market Analysis with Econometrics and Machine Learning

# 1d Hypothesis Tests

Uni Ulm

Prof. Dr. Sebastian Kranz

SoSe 2020



## Hypothesis tests: Null hypothesis

- A hypothesis test consists of a **null hypothesis** $H_0$ and a corresponding **alternative hypothesis** $H_1$ about some features of a data generating process. Examples for hypotheses for a linear regression model:
  - $H_0$: $\beta_1 = 0$, $H_1$ : $\beta_1 \neq 0$
  - $H_0$: The explanatory variable $x_k$ is exogenous, $H_1$ : $x_k$ is endogenous
  - $H_0$: The disturbance $\varepsilon$ is not auto-correlated, $H_1$: $\varepsilon$ is auto-correlated



## Example: t-test for a regression coefficient

- Consider a linear regression model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + \varepsilon$ that satisfies a multiple regression equivalent to assumptions (A1)-(A4) and the null hypothesis:

$$H_0 : \beta_k = 0$$

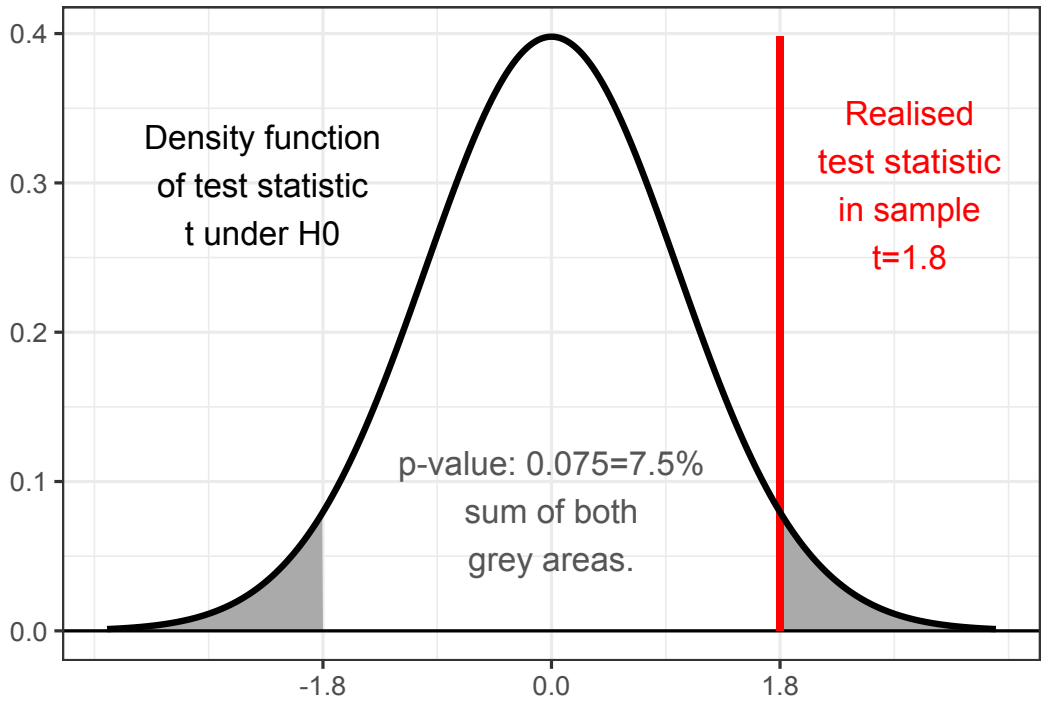
- Every hypothesis test is based on a **test statistic** that can be computed from the data. In our example, it is the following *t-value*:

$$t_k = \frac{\hat{\beta}_k}{\hat{sd}(\hat{\beta}_k)}$$

- We can also view a test statistic as a random variable. Here $t_k$ is a transformation of the random variable $\varepsilon$ and the explanatory variables.
- Key of every hypothesis test is that one knows the distribution of the test statistic if $H_0$ and all additional assumptions (here A1-A4) hold true.
  - A statistical result shows that $t_k$ is then distributed according to a $t$-distribution with $T - K - 1$ degrees of freedom if $\beta_k = 0$.

## p-value






## P-values and significance levels



- The p-value measures the probability to find the realized or more extreme test statistic if H0 is true (see plot above).
- One often considers critical levels of the p-value like 5% or 1%, which are called significance levels.
- We say we can reject the H0 at significance level $\alpha$ if the p-value is smaller than $\alpha$,
  - e.g. if we have p-value=0.043 we can reject H0 at a significance level of 5%.
- Significance levels are often marked with one or several stars ** in regression outputs.

R illustration: Run a linear regression in R and show a `summary` of the results. Explain all columns for the estimated coefficients.

## P-values and confidence intervals

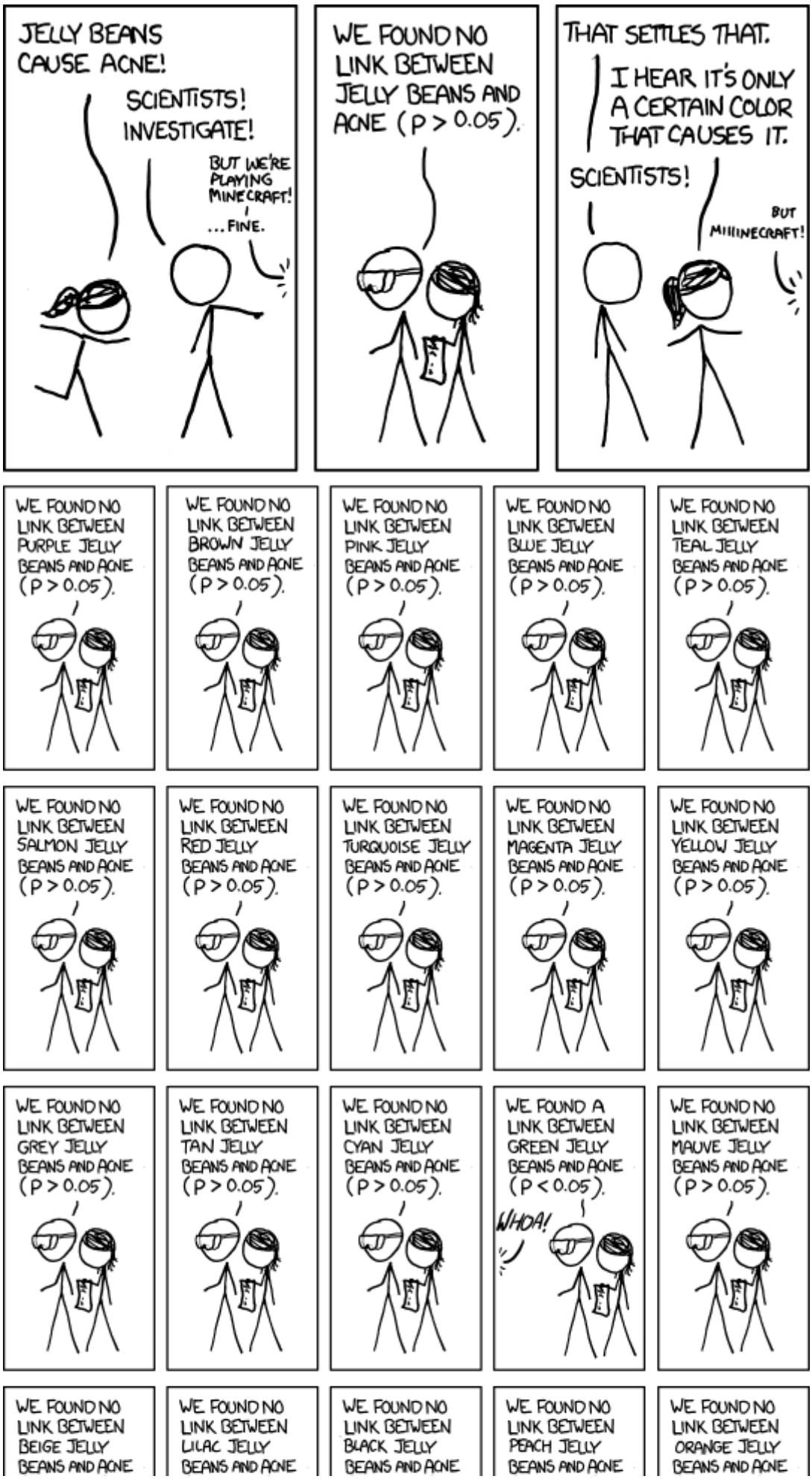
- In a linear regression the null hypothesis $\beta_k = 0$ is rejected at a significance level of 5% if and only if the 95% confidence interval around the estimate $\hat{\beta}_k$ does not include 0.

## Test questions about hypothesis testing



- Assume the null hypothesis H0 is rejected at a significance level $\alpha = 0.1\%$ (p = 0.001). Is this strong evidence that H0 is false?
- Assume the null hypothesis H0 is not rejected and we find a p-value of p = 0.999. Is this strong evidence that H0 is true?
- Assume H0 is rejected at a significance level $\alpha = 5\%$ (p = 0.05). Does it mean that the probability that H0 is true is equal to or smaller than 5%?
  - See http://xkcd.com/1132/

Source: https://xkcd.com/882/






## The problem of multiple testing and false discoveries

- Many "discoveries" in empirical sciences (e.g. whether a drug actually has a positive effects on patients) are claimed because a null hypothesis in a statistical study is rejected with a p-value below 5%.
  - But even if there is no effect, one finds a p-value below 5% in 5% of cases.
- Some observers argue that combinded with other factors this may cause many false discoveries in scientific publications.
  - See e.g. the article "Why Most Published Research Findings Are False" for different arguments why that may be the case.
  - You can also search for the term "replication crisis" for more background information.

- One problem is the so called *publication bias*. It is easier to publish new discoveries than studies that don't find significant relationships.
    - Assume, for example, two researchers who don't know each other conduct a similar behavioral sciences experiment. One researcher finds a significant effect that would be an interesting discovery, the other finds no effect. If they send it to different journals, it may well be the case that only the significant result gets finally published but not the experiment that does not find an effect. Our published scientific knowledge would then be biassed and more strongly suggest that there is an effect compared to the case that both studies would be published.
- If there are strong incentives to publish some results and lax standards at journals, an unscrupulous researcher may also try out different regression specifications (e.g. varying the set of control variables) until a significant effect is obtained. Here is a simulation about such "p-hacking":

  https://skranz.github.io/r/2018/04/09/Sample_Size_P-Values_and_Data_Mining.html

## Measures for more robust scientific insights



The scientific community is aware of the above mentioned dangers and countermeasures become increasingly important nowadays. For example:

- **Robustness checks**: Empirical studies in economics are typically required to have several robustness checks that verify that the main insights also hold for sensible alternative model specifications.
- **More replications**: The community tries to increase incentives and funding for replication studies that check whether important results indeed can be systematically replicated or were just a random finding.
- **Preregistration**: Some journals and funding organizations now require that experiments must be preregistered with a detailed plan for the statistical analysis before the experiment is run. Sometimes the experiment will be accepted for publication already based on that plan, no matter whether a new discovery is made (a p-value below 5%) or not. This shall avoid publication bias.
- Recommendations to focus less on p-values and significance. On this link is a corresponding statement by the American Statistical Association.

## Misrepresentation in business and other domains



- Many scientists are well aware of problematic statistical issues like publication bias, but I am less sure about other domains like business or politics.
- There are many ways how one can misrepresent data and empirical results, e.g. creating misleading graphs or selling relationships between two variables as causal effects even if there is a clear endogeneity problem. If somebody wants to sell an idea or product, he may unconsciously, or on purpose, do so.
- You can take a look at his website for interesting examples: https://callingbullshit.org/
- I believe that good knowledge in econometrics makes you substantially less likely to be fooled by misrepresented empirical findings.

## How to adjust for multiple testing?



- Sometimes one wants explore a lot of possible relationships and find the significant ones, e.g. to guide further studies.
    - One example are genetic studies where one want to explore whether some expressions of the over 20000 human genes are systematically linked to certain diseases.
- But how can we then correct for the fact that we tested a lot of hypotheses?

## Controlling the False Discovery Rate



- Assume we run multiple statistical tests and call all results with p-values below some critical value a *discovery*.
- The false discovery rate (FDR) is defined as the average fraction of false discoveries, i.e. the average fraction of discoveries were the null hypothesis actually was true.
- Benjamini and Hochberger (1995) proposed a simple method to guarantee that when running multiple (independent) tests, the false discovery rate is below some threshold $\delta$, e.g. $\delta = 10\%$:
    - Sort your $n$ p-values from smallest to largest. Let $p_k$ be the k-smallest p-value.
    - Find the highest $k$ such that $p_k < \delta\frac{k}{n}$.
    - If you say all results with p-values below that $p_k$ are discoveries (and signficant), then at most a share of $\delta$ of those discoveries are on average false discoveries.

## Example



Assume we have run $n = 100$ tests and want a maximum false discovery rate of $\delta = 10\%$. You see below the 6 lowest p-values:

| k | $p_k$ | $\delta\frac{k}{n}$ | $p_k \leq \delta\frac{k}{n}$ |
|---|---|---|---|
| 1 | 3.5e-05 | 0.001 | TRUE |
| 2 | 0.0015 | 0.002 | TRUE |
| 3 | 0.0033 | 0.003 | FALSE |
| 4 | 0.0035 | 0.004 | TRUE |
| 5 | 0.012 | 0.005 | FALSE |
| 6 | 0.021 | 0.006 | FALSE |
| ... | | | |

If the last column is FALSE for all further p-values, we would consider the first 4 results significant / a discovery with the Benjamini-Hochberg procedure. (Note in particular that also the third entry would be considered a discovery.)

## Diagnostic Tests



- While a t-test as discussed above is typically use to make discoveries, diagnostic tests are mainly used to check whether some assumptions of an econometric model are likely to be violated.
- For example, there are diagnostic tests to check whether disturbances $\varepsilon$ are autocorrelated, which would violate assumption A2 of the linear regression model.
- We will explore 3 diagnostic tests used for instrumentental variable estimations.

## Diagnostic tests for IV estimation



- Run an instrumental variable estimation with R and show a `summary` of the results with the option `diagnostic=TRUE`
- You see results of 3 diagnostic tests:
  - Weak instruments
  - Wu-Hausman (endogeneity of regressors)
  - Sargan (endogeneity of instruments)
- Unfortunately, currently the R-help for summary.ivreg provides almost no information what the tests do and how we should interpret these results. We will very briefly give an overview over these tests.

## Testing for weak instruments



- Consider a linear regression model of a demand function

$$q = \beta_0 + \beta_1 p + \beta_2 s + \varepsilon$$

  with endogenous prices $p$, an exogenous explanatory variables $s$.
- We also shall have two excluded instruments $z_1$ and $z_2$, e.g. two factors that influence costs and thereby prices.
- The weak instruments problem means that if the instruments $z_1$ and $z_2$ are only weakly correlated with $p$ the IV estimator can become considerably biassed (and imprecise) for small sample size $T$.
- The test for weak instruments shown in R tests the null hypothesis that in the first stage regression of the two stage least squares procedure

$$p = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 s + \eta$$

  the coefficients of the excluded instruments are zero, i.e. here:

  H0: $\gamma_1 = \gamma_2 = 0$
- This is a so called F-test and its test statistic is called F-statistic.
- Rule-of-thumb: Staiger and Stock (1997) suggested declaring instruments to be weak if the F-statistic is smaller than 10 (not looking at the p-value), Stock and Yogo (2005) provide much more details.

## Wu-Hausman test for endogenous regressors



- Consider a linear regression model of a demand function

$$q = \beta_0 + \beta_1 p + \beta_2 s + \varepsilon$$

  for which we don't know if prices $p$ are endogenous or exogenous.
- If we have valid instruments $z$ for a possibly endogenous variable $p$, the *Wu-Hausman test* allows to test whether $p$ is indeed endogenous.
- The null hypothesis of the Wu-Hausman test is that all explanatory variables of a regression are exogenous
  - i.e. low p-values of the Wu-Hausman test suggest an endogenous variable.

- The Sargan test is a test with the Null hypothesis that all instruments are exogenous.
- The Sargan test can only be applied if we have at least one more excluded instrument than endogenous variable.
- If the Sargan test is rejected (low p-value), it suggests that at least one instrument is endogenous.
- But: If the Sargan test is not rejected we do **not** have strong proof that all instruments are indeed exogenous, e.g. the Sargan test may well fail to detect if all instruments are endogenous.
    - This means not being rejected by the Sargan test can be interpreted as a neccessary condition for exogenous instruments but not a sufficient one. Most important remains the economic reasoning behind the selection of the instruments.