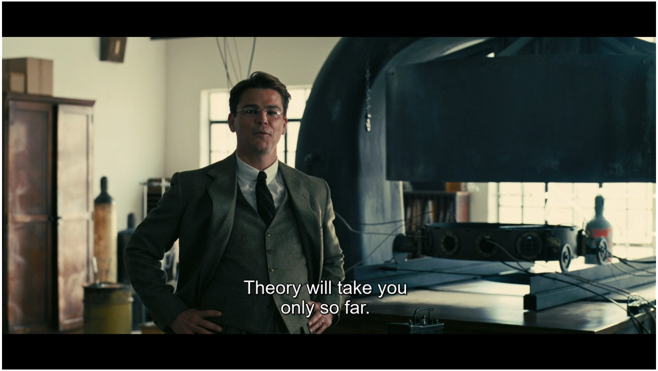
# Hypothesis Testing

Ye Wang
University of North Carolina at Chapel Hill

*Mathematics and Statistics For Political Research*
*POLI783*

# From theory to empirics

- Why do we need empirical works in social science?




- It is critical to test whether our theory makes sense in reality.
- Otherwise, it is better to revise the theory.

# From theory to empirics

- From our theory, we can often derive testable implications.
- E.g., the choice to turn out in elections is motivated by "civic duty" (Riker and Ordeshook, 1968).
- If we increase the sense of civic duty among voters, turnout rate will be higher (Gerber and Green, 2000).
- To test this implication, they implemented a randomized experiment and estimated the effect of get-out-to-vote (GOTV) messages on turnout.
- Let $\tau_i$ be the effect of civic duty on voter $i$.
- The theory suggests that $\tau_i > 0$, thus $\tau = \mathbb{E}[\tau_i] > 0$, or at least $\tau \neq 0$.
- This is a hypothesis: a statement regarding functionals of the DGP (estimands).
- They obtained an effect estimate of 8.5% and a standard error estimate of 2.6%.
- How do we know whether the results support the theory?

# Hypothesis testing

- A hypothesis is about an estimand.
- But all we have are the data, $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_N) \subset \mathbb{R}^{N \times P}$, and the estimates.
- The process to evaluate the hypothesis using estimates is known as hypothesis testing.
- It builds on the idea of falsification.
- We can never verify that all swans are white.
- But with one black swan, we can reject this hypothesis.
- Therefore, we usually start from the opposite of the hypothesis implied by our theory and try to reject it.
- It resembles the proof by contradiction.

# Classification of hypotheses

- If we believe that $\tau \neq 0$, we should test the hypothesis that $\tau = 0$.
- This is known as the null hypothesis and denoted as $H_0$.
- Accordingly, $\tau \neq 0$ is known as the alternative hypothesis and denoted as $H_1$.
- What are the null and alternative hypotheses if we believe that $\tau > 0$?
- We sometimes refer to $\tau \neq 0$ as a two-sided alternative hypothesis and $\tau > 0$ as a one-sided alternative hypothesis.
- Hypotheses such as $\tau = 0$ are defined with averages or expectations and sometimes referred to as a weak null.
- There are also sharp null hypotheses such as $\tau_i = 0$ for any $i$.
- We can test multiple hypotheses together, which is known as a joint hypothesis.
- 0 here can be replaced by any value $\tau_0$.

# Rejection region

- We want to determine whether $H_0$ should be rejected using our data $\mathbf{O}$.
- Ideally, we would construct a "rejection region" $R \subset \mathbb{R}^{N \times P}$ and reject $H_0$ if and only if $\mathbf{O} \in R$.
- In practice, we choose a "test statistic" $T$ that maps $\mathbf{O}$ to a real number.
- A common choice is the $t$-statistic:

$$T(\tau) = t_N(\tau) = \frac{\hat{\tau}_N - \tau}{\hat{\sigma}_{\hat{\tau}_N}}.$$

- Then, the rejection region becomes an interval, and we reject $H_0$ if and only if $|T(\tau_0)| > c$ (two-sided test) or $T(\tau_0) > c$ (one-sided test).
- Here $c$ is known as the critical value, and $R = \{t : |t| > c\}$ or $\{t : t > c\}$.
- A test consists of the test statistic and the rejection region.

# Type I and type II errors

- Based on the test's result and whether $H_0$ is true, there are four outcomes:

| | $H_0$ True | $H_0$ False |
|---|---|---|
| Retain $H_0$ | Awesome | Type II error |
| Reject $H_0$ | Type I error | Awesome |

- **Type I error**: rejecting the null hypothesis when it is in fact true (false positive).
- **Type II error**: not rejecting the null hypothesis when it is false (false negative).
- Type I error is usually more severe in reality.
- Announcing pregnancy when you are not pregnant vs. failing to detect pregnancy when you are pregnant.

# Features of a test

- We want to avoid both types of errors in practice.
- Our test should reject the null when it is false and retain it otherwise.
- We can evaluate a test's quality via its probability of committing either type of error.
- The size of a test is defined as the probability of committing the type I error:

$$\pi(\tau_0) = \mathbb{P}\left(\text{Reject } H_0 \mid \tau = \tau_0\right).$$

- For a two-sided test, $\pi(\tau_0) = \mathbb{P}\left(|T(\tau)| > c \mid \tau = \tau_0\right)$.
- The power of a test is defined as

$$\pi(\tau_1) = \mathbb{P}\left(\text{Reject } H_0 \mid \tau = \tau_1\right).$$

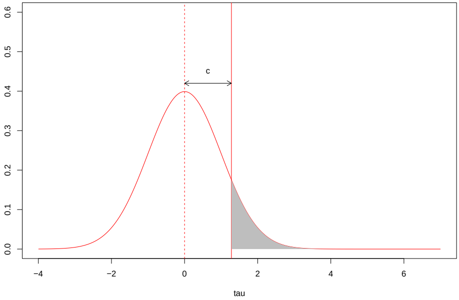
- A test's power equals 1 minus the probability of committing the type II error.
- We want to minimize the size while maximizing the power.

# Features of a test

- These two goals often conflict with each other in practice.
- Consider the $t$-statistic; we know that

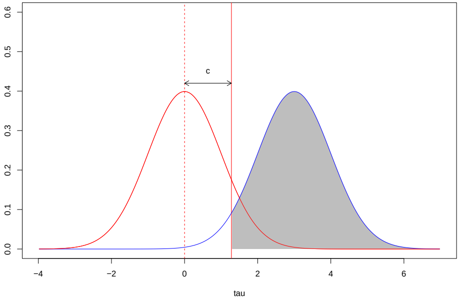
$$t_N = \frac{\hat{\tau}_N - \tau}{\hat{\sigma}_{\hat{\tau}_N}} \xrightarrow{d} \mathcal{N}(0, 1) .$$

- Suppose $\tau_0 = 0$, then we reject $H_0$ when $\frac{\hat{\tau}_N}{\hat{\sigma}_{\hat{\tau}_N}} > c$.



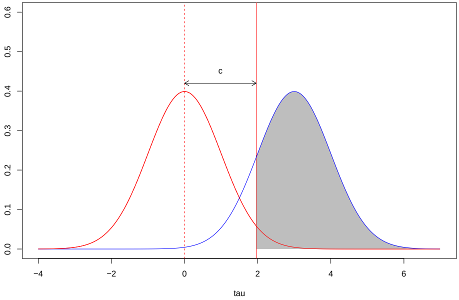
# Features of a test

- If $\tau_1 = 3$ under the alternative hypothesis, then the test's power is the shaded region's area.



# Features of a test

- When $c$ is larger, the test's size becomes smaller and so does the power.



# Significance level

- As we are more concerned with type I error than type II error, the practice is just to fix the size and let the power be.
- We will choose a "significance level" $\alpha$ and require that the size is below this level.
- Then, we solve the critical value for this requirement to be satisfied.
- E.g., let $\alpha = 0.05$, then the probability of committing type I error will be below 5%.
- For the one-sided test based on $t$-statistic, it means that

$$\pi(\tau_0) = \mathbb{P}\left(t_N > c \mid \tau = \tau_0\right) = \mathbb{P}_{\tau_0}\left(\frac{\hat{\tau}_N - \tau_0}{\hat{\sigma}_{\hat{\tau}_N}} > c\right) \leq 0.05.$$

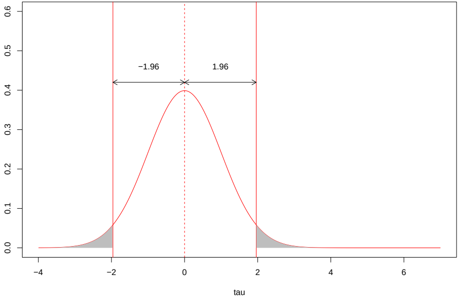
- As $t_N$ can be approximated by the standard normal distribution, we just need to find the 95% quantile for $\mathcal{N}(0,1)$, which is 1.645.

# Significance level

- If $t_N > 1.645$, we can reject $H_0$ with a lower-than-5% probability to commit type I error.
- For a two-sided test, we want to find $c$ such that

$$\mathbb{P}_{\tau_0}\left(\left|\frac{\hat{\tau}_N - \tau_0}{\hat{\sigma}_{\hat{\tau}_N}}\right| > c\right) \leq 0.05.$$

- As the standard normal distribution is symmetric, we can set $c$ as its 97.5% quantile, 1.96.



# Significance level

- Two-sided tests are more common in practice as they are more conservative.
- A common approach is to compute the p-value:

$$p = 1 - \Phi\left(\frac{\hat{\tau}_N - \tau_0}{\hat{\sigma}_{\hat{\tau}_N}}\right) \text{ or } p = 2\left(1 - \Phi\left(\left|\frac{\hat{\tau}_N - \tau_0}{\hat{\sigma}_{\hat{\tau}_N}}\right|\right)\right)$$

  and compare it with the significance level.
- E.g., if $p = 0.037$, the estimate is significant at the 5% level.
- There is nothing magical about 5%.
- People also use 10% or 1%.
- It is just a convention in social science.
- In CERN, the widely-accepted $\alpha$ is $\frac{1}{1{,}750{,}000}$.

# The p-value

- The p-value is computed from data and thus a r.v.
- We can show that $p \sim Unif(0,1)$:

$$\begin{aligned}\mathbb{P}(p \leq x) =& \mathbb{P}\left(1-\Phi\left(t_N\right) \leq x\right) = 1-\mathbb{P}\left(\Phi\left(t_N\right) \leq 1-x\right) \\ =& 1-\mathbb{P}\left(t_N \leq \Phi^{-1}\left(1-x\right)\right) = 1-\left(1-x\right) = x.\end{aligned}$$

- The distribution does not vary with $N$ under the null.
- But if the alternative holds, the p-value converges to zero as $N$ grows, and the test's power increases.
- Under $H_0$, both $\hat{\tau}_N - \tau_0$ and $\hat{\sigma}_{\hat{\tau}_N}$ converge to zero.
- Under $H_1$, the former converges to $\tau_1 - \tau_0$, while the latter still converges to zero.
- No need to lower the significance level when $N$ is larger.
- The p-value does not tell you the effect's magnitude or the probability for either hypothesis to be true.

# Confidence intervals

- Equivalently, we can construct confidence intervals (CI) at any significance level.
- From the definition of the t-statistic, we have

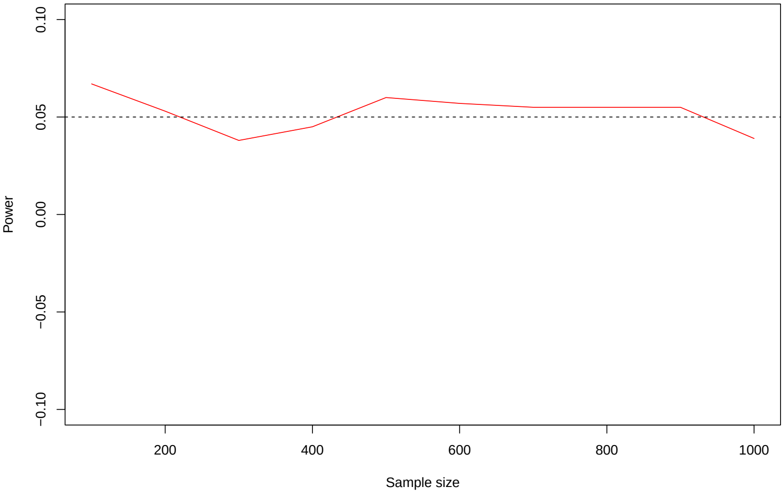
$$
\begin{aligned}
&\mathbb{P}_{\tau_0}\left(\left|\frac{\hat{\tau}_N-\tau_0}{\hat{\sigma}_{\hat{\tau}_N}}\right|>c\right)=1-\mathbb{P}_{\tau_0}\left(\left|\frac{\hat{\tau}_N-\tau_0}{\hat{\sigma}_{\hat{\tau}_N}}\right|\leq c\right)\\
=&1-\mathbb{P}_{\tau_0}\left(-c\hat{\sigma}_{\hat{\tau}_N}\leq\hat{\tau}_N-\tau_0\leq c\hat{\sigma}_{\hat{\tau}_N}\right)\\
=&1-\mathbb{P}_{\tau_0}\left(\hat{\tau}_N-c\hat{\sigma}_{\hat{\tau}_N}\leq\tau_0\leq\hat{\tau}_N+c\hat{\sigma}_{\hat{\tau}_N}\right)\leq\alpha.
\end{aligned}
$$

- Therefore, the probability that the interval $[\hat{\tau}_N-c\hat{\sigma}_{\hat{\tau}_N},\hat{\tau}_N+c\hat{\sigma}_{\hat{\tau}_N}]$ covers $\tau_0$ is larger than $1-\alpha$.
- If $\alpha=0.05$, then we can set $c=1.96$.
- If $\hat{\tau}_N-c\hat{\sigma}_{\hat{\tau}_N}>0$ or $\hat{\tau}_N+c\hat{\sigma}_{\hat{\tau}_N}<0$, then we can reject $H_0$ at the level of 5%.
- CIs can be asymmetric, but this is uncommon in practice.

# Summary

- From the frequentist perspective, $\tau_0$ is a functional and thus a fixed quantity.
- The significance level $\alpha$ is chosen by researchers to control the test's size hence also a constant.
- Both $\hat{\tau}_N$ and $\hat{\sigma}_{\hat{\tau}_N}$ are random variables as they are estimated from data.
- Therefore, any test statistic is a random variable.
- The associated p-value and confidence intervals are also random variables.
- If the randomly generated p-value is smaller than the fixed significance level, we say the estimate is statistically significant at this level.
- The 95% confidence interval should cover $\tau_0$ in 95% of repeated samples.

# Summary



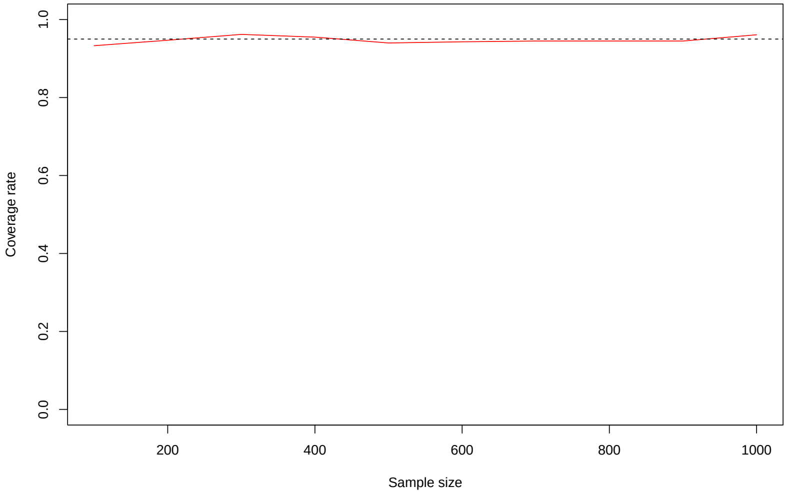
# Coverage rate

- We define the coverage rate of a confidence interval $[\hat{\tau}_L, \hat{\tau}_U]$ as

$$\mathbb{P}_{\tau_0}\left(\hat{\tau}_L \leq \tau_0 \leq \hat{\tau}_U\right).$$

- Does the 95% CI we saw have a coverage rate of at least 95%?
- The answer is sometimes negative, especially when the estimator is complex, $N$ is small, and the variance is heteroskedestic.
- There are several approximations: we approximate $\sigma_{\hat{\tau}_N}$ with $\hat{\sigma}_{\hat{\tau}_N}$ and the sampling distribution's critical values with the normal distribution's critical values.

# Coverage rate



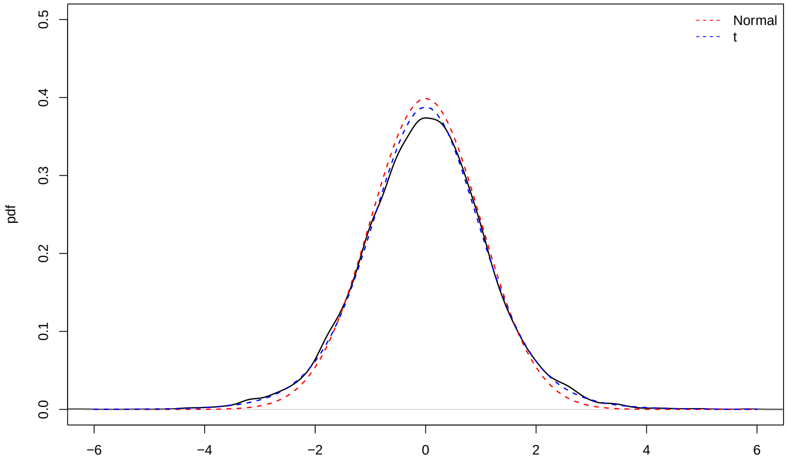
# The exact test

- This problem is less severe if the data are drawn from a normal distribution.
- We have proved that for the sample average,

$$\frac{\bar{X}_N - \mu}{\sqrt{s_N^2/N}} \sim t_{N-1}.$$

- We have an exact test rather than an asymptotic approximation by using critical values from the student-t distribution.
- Remember that the student-t distribution has a fatter tail: the critical values are larger than those from the standard normal distribution.
- It is a more conservative approach in small samples.

# The exact test



# The Neyman-Pearson approach (*)

- In theory, we choose the rejection region to maximize the test's power while requiring that its size must be below a level:

$$R^* = \arg \max_{R \subset \mathbb{R}^{N \times P}} \pi(\tau_1), \text{ s.t. } \pi(\tau_0) \leq \alpha.$$

- This is known as the Neyman-Pearson approach.
- But this is unfeasible as we usually do not know $\tau_1$.
- For a simple alternative hypothesis $\tau = \tau_1$, Neyman and Pearson show that the problem has a solution.
- We find $c$ such that

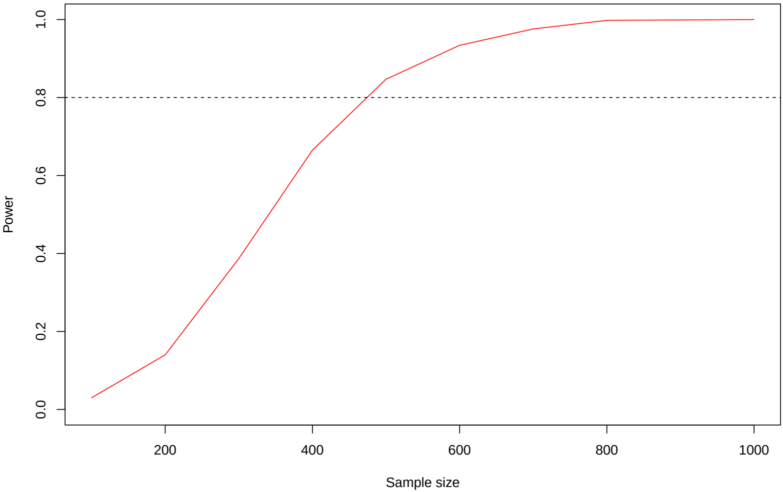
$$\mathbb{P}\left(f_{\tau_1}(\mathbf{O}) > c f_{\tau_0}(\mathbf{O}) \mid \tau = \tau_0\right) = \alpha,$$

  where $f_\tau(\mathbf{O})$ is the data's joint p.d.f. under $\tau$.
- Then, we reject $H_0$ if and only if $f_{\tau_1}(\mathbf{O}) > c f_{\tau_0}(\mathbf{O})$.
- It leads to the t-statistic when $f_\tau(\cdot)$ is normal.
- This is known as a "uniformly most powerful" (UMP) test.

# Power analysis

- In practice, we can only fix the test's size as its power depends on $\tau_1$.
- It is still necessary to evaluate the test's power under different $\tau_1$s.
- This process is known as power analysis.
- What is the probability for a test to reject $H_0$ if $\tau = \tau_1$ and $\sigma/\sqrt{N}$ is known?
- This can be estimated via simulation.
- We can change any of these components to examine how the power varies.

# Power analysis: simulation



# Multiple testing

- In practice, we sometimes want to test the joint null hypothesis:

$$H_0 : \tau_1 = \tau_{10}, \tau_2 = \tau_{20}, \ldots, \tau_K = \tau_{K0}.$$

- E.g., we are often interested in the effect of a policy on multiple outcomes.
- If these outcomes are independent and we use the t-test with $c = 1.96$, then under the joint null,

$$\begin{aligned} & \mathbb{P}_{\tau_0}\left(|t_{1N}| \leq 1.96, |t_{2N}| \leq 1.96, \ldots, |t_{KN}| \leq 1.96\right) \\ = & \prod_{k=1}^{K} \mathbb{P}_{\tau_0}\left(|t_{kN}| \leq 1.96\right) \geq 0.95^K. \end{aligned}$$

- If $K = 20$, then $0.95^K = 0.36$, and the probability for at least one hypothesis to be rejected is $1 - 0.36 = 0.64$.
- We are very likely to commit the type I error.

# Multiple testing

- The probability of rejecting at least one hypothesis under the joint null is known as the family-wise error rate (FWER).
- One way to ensure that the FWER is below 0.05 is to use the Bonferroni correction.
- We set the significance level as $\tilde{\alpha} = \alpha/K$.
- In this case,

$$
\begin{aligned}
& 1 - \mathbb{P}_{\tau_0}\left(|t_{1N}| \le 1.96, |t_{2N}| \le 1.96, \ldots, |t_{KN}| \le 1.96\right) \\
\le & \sum_{k=1}^{K} \mathbb{P}_{\tau_0}\left(|t_{kN}| > 1.96\right) \le \sum_{k=1}^{K} \frac{\alpha}{K} = \alpha.
\end{aligned}
$$

- In reality, the Bonferroni correction can be too conservative.
- E.g., even when the outcomes are independent, the FWER equals $1 - \left(1 - \frac{\alpha}{K}\right)^K \approx \alpha - \frac{\alpha^2}{2K} < \alpha$.
- The FWER is smaller when the outcomes are dependent, therefore alternatives may be needed.