Hypothesis Testing

A Motivating Example

- Between 1960 and 1980, there were many lawsuits in the South claiming racial bias in jury selection.
- Here's some made up[*] (but similar) supporting data:
  - 50% of citizens in the local area are African American
  - On an 80 person panel, only 4 were African American
- Can this outcome be explained as the result of pure chance?
- If $X \sim \text{Binomial}(n = 80, p = 0.5)$, then $P[X = 4] \approx P[X \leq 4] \approx 1 \times 10^{-18}$
- *N.B.*: Statistics can never *prove* anything.
  Still, this outcome is extremely unlikely to be the result of pure chance!

[*]This example was borrowed from *The Cartoon Guide to Statistics*.

Hypothesis Testing, the basics

- A hypothesis test is designed to test whether observed data is “as expected”, as described by a statistical model.
  - Are the colors in a bag of M&Ms distributed as expected?
  - Did the proportion of the U.S. adult population who support environmental regulations change in the past year, the past decade, the past century, etc.?
  - Are there fewer COVID cases among the vaccinated?
- A test statistic is a measure of the observed data.
- A hypothesis test then compares the test statistics to what was expected, by comparing the probability of the test statistic with a significance level, and rejects the model of what was expected if this probability is sufficiently low.
- A significance level (α) is a cutoff, determined in advance, below which we will declare that we have observed something other than what was expected.

Hypothesis Testing, in more detail

- Step 1: Formulate a null and an alternative hypothesis
  - The null hypothesis is a claim that data are distributed in some way (e.g., $B(80, 0.5)$)
  - An alternative hypothesis is a claim that data are distributed in some other way (e.g., $p < 0.5$)
  - The null is so-called because it is usually a claim about no significant effect or difference, and it is often something we suspect the data will disprove.
- Step 2: Compute a test statistic
  - A test statistic summarizes the observed data.
- Step 3: Find the *p*-value of the test statistic
  - What is the probability of observing this value of the test statistic, under the null hypothesis?
  - A *p*-value measures the extent to which an observed sample of data agrees with an assumed probability model (i.e., the distribution under the null hypothesis).
- Step 4: Determine whether the test statistic is significant
  - If the *p*-value $< \alpha$, then the test is deemed significant, and the null hypothesis is rejected.

Example Test Statistics

$$z = \frac{\overline{x} - \mu_0}{(\sigma/\sqrt{n})}$$

$$t = \frac{\overline{x} - \mu_0}{(s/\sqrt{n})},$$

$$df = n - 1$$

$$\chi^2 = \sum^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Back to the Example

- Numerator: $p_{hat} - p = (4/80) - 0.5 = 0.05 - 0.5 = -0.45$
  - By subtracting $p = 0.5$, we are assuming the null hypothesis is $p = 0.5$.
- Denominator: Standard Error
  - $\mathrm{Var}[p_{hat}] = (0.5)(1 - 0.5)/80 = 0.003125$
  - The standard error is the square root of this variance: $\sqrt{0.003125} = 0.056$
- z-statistic: $-0.45/0.056 \approx -8.0$
  - That’s a whole lot of standard deviations below the mean!
- If the null hypothesis were true, the probability of observing this value of our test statistic is essentially 0.
- We reject the null hypothesis and search for alternative explanations.

Two Sides of the Same Coin

Hypothesis testing and confidence intervals are two sides of the same coin.

- The 95% CI is $Pr[p_{hat} - z_{lo}\sigma_{hat} \leq \mu \leq p_{hat} + z_{hi}\sigma_{hat}] = .95$
  - Lower Bound: 4/80 + (-1.96)(0.056) = -0.06
  - Upper Bound: 4/80 + (1.96)(0.056) = 0.15
- This interval does not contain 0.5, the null hypothesis.
- We reject the null hypothesis and search for alternative explanations.

All the Steps in Hypothesis Testing

- Step 0: Set a significance level (α)
- Step 1: Formulate null and alternative hypotheses
- Step 2: Compute a test statistic, assuming the null hypothesis
- Step 3: Find the *p*-value of the test statistic, assuming the null hypothesis
- Step 4: Compare the *p*-value to α
  - If the *p*-value $< \alpha$, then the test is deemed significant, and the null hypothesis is rejected
  - Otherwise, the test is insignificant, and the null hypothesis cannot be rejected

Language of Hypothesis Testing

- Critical value: the value of the test statistic at which the null hypothesis is rejected, given the significance level (α)
- Rejection region: the set of values of the test statistic for which the null hypothesis is rejected. (The non-rejection region is defined analogously.)

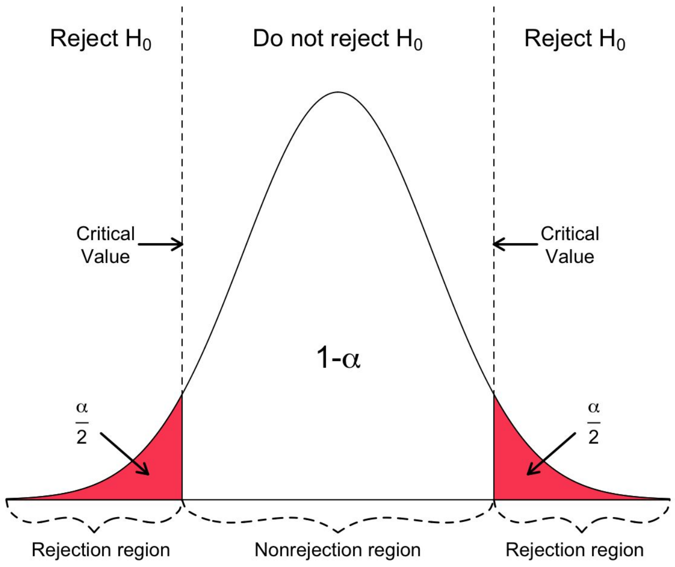


Image Source

One-sided vs. Two-sided tests

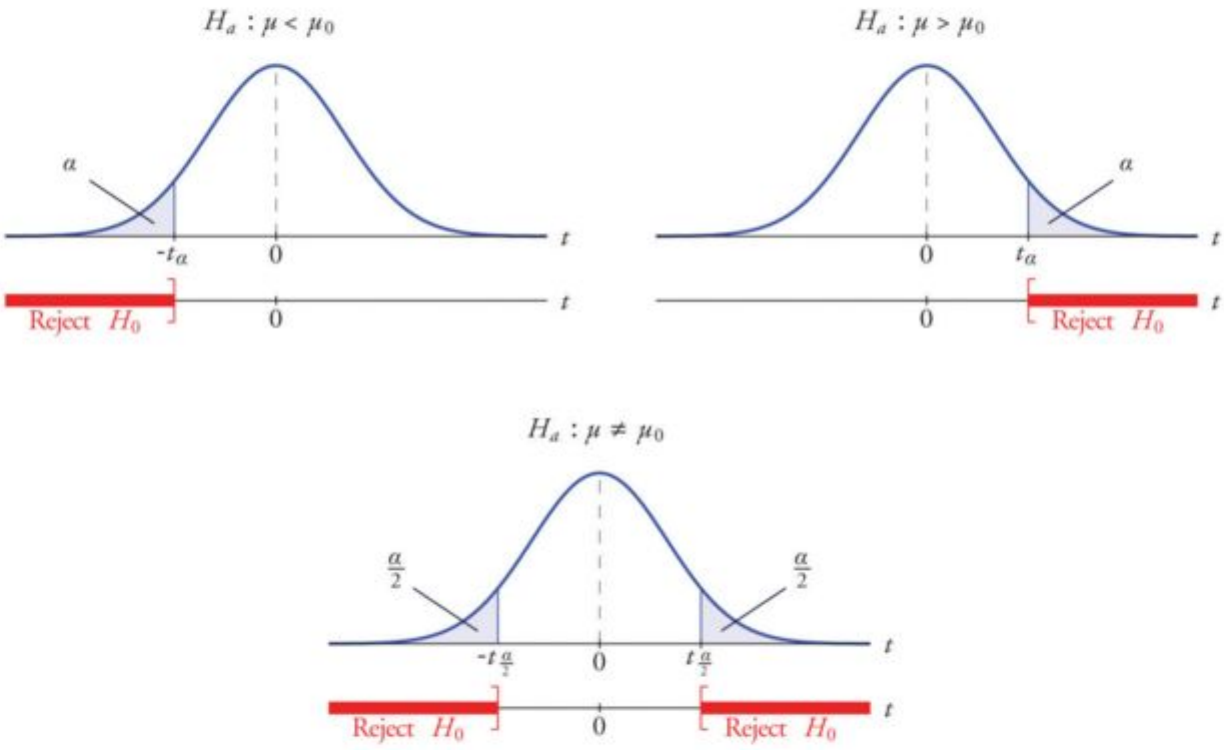


Image Source

One-sided vs. Two-sided tests

- Let's test whether a coin is biased. The null hypothesis is that the coin is fair.
- Possible alternative hypotheses include: $p_H > p_T$, $p_H < p_T$, and $p_H \neq p_T$.
- Assume all heads are observed
  - $p_H > p_T$: The null is rejected.
  - $p_H < p_T$: The null is *not* rejected.
  - $p_H \neq p_T$: The null is rejected.
- Assume all tails are observed
  - $p_H > p_T$: The null is *not* rejected.
  - $p_H < p_T$: The null is rejected.
  - $p_H \neq p_T$: The null is rejected.

John Snow’s Grand Experiment, Revisited Again

Data collected by John Snow

Supply Area	# of Houses	Cholera Deaths	Deaths/10,000 Houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Last time: Confidence Intervals

- For S&V:
  - P_{hat} = 1263/40046 ≈ 0.0315
  - Var[p_{hat}] = (0.0315)(1 - 0.0315) / 40046 = 7.62 x 10e-7
  - The standard error is the square root of the variance: SE[p_{hat}] = √7.62 x 10e-7 = 0.00087
  - The CI at the 95% level is: [0.0315 - (1.96)(0.00087), 0.0315 + (1.96)(0.00087)] = [0.03, 0.033]

- For Lambeth:
  - P_{hat} = 98/26107 ≈ 0.00375
  - Var[p_{hat}] = (0.00375)(1 - 0.00375) / 26107 = 1.43 x 10e-7
  - The standard error is the square root of the variance: SE[p_{hat}] = √1.43 x 10e-7 = 0.000378
  - The CI at the 95% level is: 0.00375 - (1.96)(0.000378), 0.00375 + (1.96)(0.000378)] = [0.003, 0.0044]

Step 1: Formulate the Hypotheses

- Our null hypothesis is that the proportion of people that died in the S&V area is equal to the proportion of people who died in Lambeth.
  - $p_{S\&V} = p_L$: i.e., $p_{NULL} = 0$
- Our alternative hypothesis is that the proportion of people that died in the S&V area is greater than the proportion of people who died in Lambeth.
  - $p_{S\&V} > p_L$

Step 1: Formulate the Hypotheses

- Our null hypothesis is that the proportion of people that died in the S&V area is equal to the proportion of people who died in Lambeth.
  - $p_{S\&V} = p_L$: i.e., $p_{NULL} = 0$
- Our alternative hypothesis is that the proportion of people that died in the S&V area is greater than the proportion of people who died in Lambeth.
  - $p_{S\&V} > p_L$
- Other plausible alternative hypotheses include:
  - $p_{S\&V} < p_L$
  - $p_{S\wedge V} \neq p_L$: i.e., $p_{S\&V} > p_L$ or $p_{S\&V} < p_L$
- Our choice tests whether there is a difference in one or both directions.

Step 2: Calculate the Test Statistic

- *z*-statistic for the difference between two sample proportions
- Numerator: $(p_{S\&V} - p_L) - p_{NULL}$ = (0.0315 - 0.00375) - 0 = 0.02775
    - By subtracting 0, we are assuming the null hypothesis (i.e., it is our baseline).
- Denominator: Standard Error
    - 40046 people lived in S&V
    - 26107 people lived in Lambert
    - $Var[p_{S\&V} - p_L]$ = (0.0315)(1 - 0.0315)/40046 + (0.00375)(1 - 0.00375)/26107 = 9.05 x 10e-7
    - The standard error is the square root of this variance: √9.05 x 10e-7 = 0.00095
- *z*-statistic: 0.02775 / 0.00095 = 29.21

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Step 3: Calculate the *p*-value

- What is the probability of observing this value of the test statistic?

```
> pnorm(29.21, lower.tail = FALSE)
[1] 7.237886e-188
```

- Under the null hypothesis, there was a very low probability that an equal proportion of people died in both areas.

- This means one of two things:
  - We witnessed something incredibly rare.
  - The assumption that the null hypothesis is true is incorrect.

Steps 0 and 4: Hypothesis Testing

- Typically, using expert knowledge, the researcher sets a benchmark threshold (α-level) before running the test.
- Often, the threshold is 5% (corresponding to a 95% confidence interval).
- If the *p*-value is below this threshold, then the test is deemed significant, and the null hypothesis is rejected. A search for alternative explanations ensues.

Intuitively, since $29.21 > 1.645$*, we reject the null hypothesis at the $\alpha = 0.05$ level.
Likewise, since 7e-188 < 0.05, we reject the null hypothesis at the $\alpha = 0.05$ level.

*Recall we are performing a one-sided test!

Two Sides of the Same Coin

Hypothesis testing and confidence intervals are two sides of the same coin.

- The 95% CI is $Pr[p_{hat} - z_{lo}\sigma_{hat} \leq \mu \leq p_{hat} + z_{hi}\sigma_{hat}] = .95$
  - Lower Bound: 0.02775 + (-1.96)(9.05 x 10e-7) = 0.02773
  - Upper Bound: 0.02775 + (1.96)(9.05 x 10e-7) = 0.02777
- This interval does not contain 0, the null hypothesis.
- We reject the null hypothesis and search for alternative explanations.

Step 1: Formulate the Hypotheses

- Our null hypothesis is that the proportion of people that died in the S&V area is equal to the proportion of people who died in Lambeth.
  - $p_{S\&V} = p_L$: i.e., $p_{NULL} = 0$
- Our alternative hypothesis is that the proportion of people that died in the S&V area is greater than the proportion of people who died in Lambeth.
  - $p_{S\&V} \neq p_L$

Step 3: Calculate the *p*-value

- What is the probability of observing this value of the test statistic?

```
> 2 * pnorm(29.21)
[1] 1
```

- Under the null hypothesis, there was a very low probability that an equal proportion of people died in both areas.
- This means one of two things:
  - We witnessed something incredibly rare.
  - The assumption that the null hypothesis is true is incorrect.

Chi-Squared Distribution

Reference: Inferential Thinking

Jurors in Alemeda County

- A total of 1453 people reported for jury duty in Alameda County in Northern California between 2009 and 2010.

	Asian	Black	Latinx	White	Other
Population	15%	18%	12%	54%	1%
Jurors	26%	8%	8%	54%	4%

- Were juries representative of the population from which they were drawn?

Exploratory Data Analysis

	Asian	Black	Latinx	White	Other
Population	15%	18%	12%	54%	1%
Jurors	26%	8%	8%	54%	4%

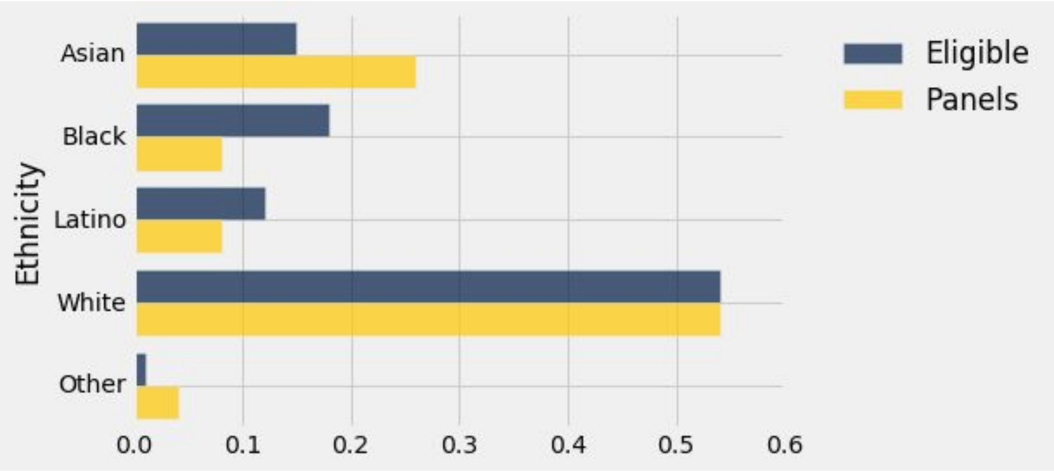


Image Source

Pearson's Chi-squared Test

- Tests whether the difference between the observed and expected frequencies of multiple categories is statistically significant.
- The multinomial distribution generalizes the binomial.
  - The binomial models the counts of flipping a coin n times
  - The multinomial models the counts of rolling a k-sided die n times
  - Bernoulli : binomial as categorical : multinomial
- Null hypothesis: Juror distribution is consistent with that of the population. I.e., jurors are distributed according to a multinomial with probabilities (0.15, 0.18, 0.12, 0.54, 0.01).
- The chi-squared test statistic measures the difference between the observed and the expected distributions.

Chi-squared Test Statistic

The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

where

- χ^2 Pearson's chi-squared test statistic
- O_i the number of observations of type *i*
- N the total number of observations
- $E_i = Np_i$ the expected (theoretical) frequency of type *i*, asserted by the null hypothesis, namely that the proportion of type *i* in the population is p_i
- n the number of types

Chi-squared Test Statistic, cont’d

- For Asians: $(26\% - 15\%)^2 / 15\% = 0.0807$
- For Blacks: $(8\% - 18\%)^2 / 18\% = 0.0556$
- For Latinx: $(8\% - 12\%)^2 / 12\% = 0.0133$
- For Whites: $(54\% - 54\%)^2 / 54\% = 0$
- For Others: $(1\% - 4\%)^2 / 4\% = 0.0225$

The chi-square test statistic is thus:
1453 (0.0807 + 0.0556 + 0.0133 + 0.0225) = 250

Chi-squared Distribution

- The distribution of the sum of the squares of *k* independent standard normal random variables
- The chi-square distribution is parameterized by degrees of freedom

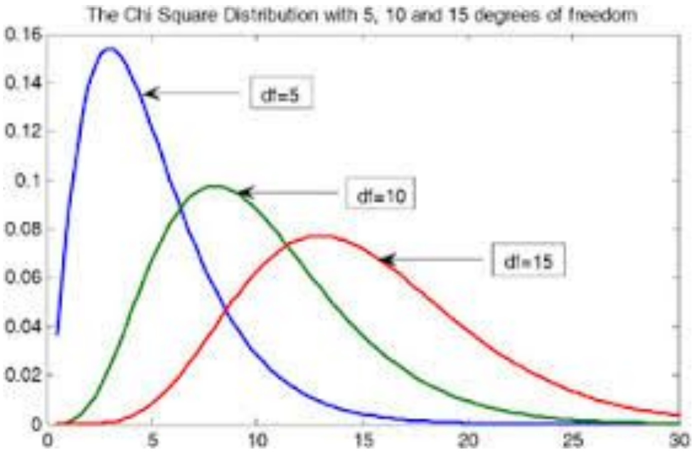


Image Source

Conclusion

- Choose α = 95%. Since there are 5 races, there are 4 degrees of freedom.
    - `qchisq(.95, df = 4)`
      `[1] 9.487729`

- Since 250 > 9.487729, we reject the null hypothesis
- Likewise, the *p*-value is essentially 0:
    - `pchisq(250, df = 4, lower.tail = FALSE) = 6.50969e-53`

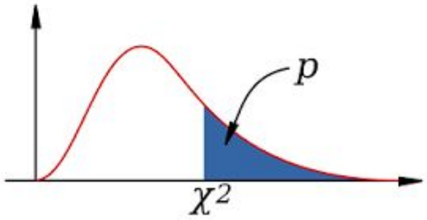


Image Source

- We reject the null hypothesis:
  Juries were not racially representative in Alameda County in 2009 and 2010.

Errors

“Innocent until proven guilty”

- Hypothesis testing is a statistical implementation of this maxim
- Null hypothesis: the defendant is innocent
- Alternative hypothesis: the defendant is guilty
  - A type 1 error (false positive) occurs when we put an innocent person in jail
  - A type 2 error (false negative) occurs when we do not jail a guilty person
- Another example:
  - Type 1 error: false alarm (fire alarm when there is no fire)
  - Type 2 error: fire but no fire alarm
- In sum:
  - Type 1 error: we reject the null hypothesis when we should not
  - Type 2 error: we do not reject the null hypothesis when we should

Type I vs. Type II Errors

- Cancer screening
  - Null hypothesis: no cancer
  - Type I: cancer suspected where there is none—not good, but not terrible
  - Type II: cancer goes undetected—very very bad
- Err on the side of type I errors
  - Make it easy to reject the null, even when we should not
  - Choose higher significance level (i.e., higher α)

Type I vs. Type II Errors

- Spam Filters
  - Null hypothesis: an email is legitimate
  - Type I: filter a legitimate email—could be very bad
  - Type II: don't filter spam—not so bad
- Err on the side of type II errors
  - Make it hard to reject the null, even when we should
  - Choose lower significance level (i.e., lower α)

Type I vs. Type II Errors

- Suspected terrorists
  - Null hypothesis: person is not a terrorist
  - Type I: send an innocent person to Guantánamo Bay
  - Type II: let a terrorist (who intends to commit mass murder) free
- US has erred on the side of type I errors, which explains why people are often held at Guantánamo Bay without a fair trial

Statistical Power

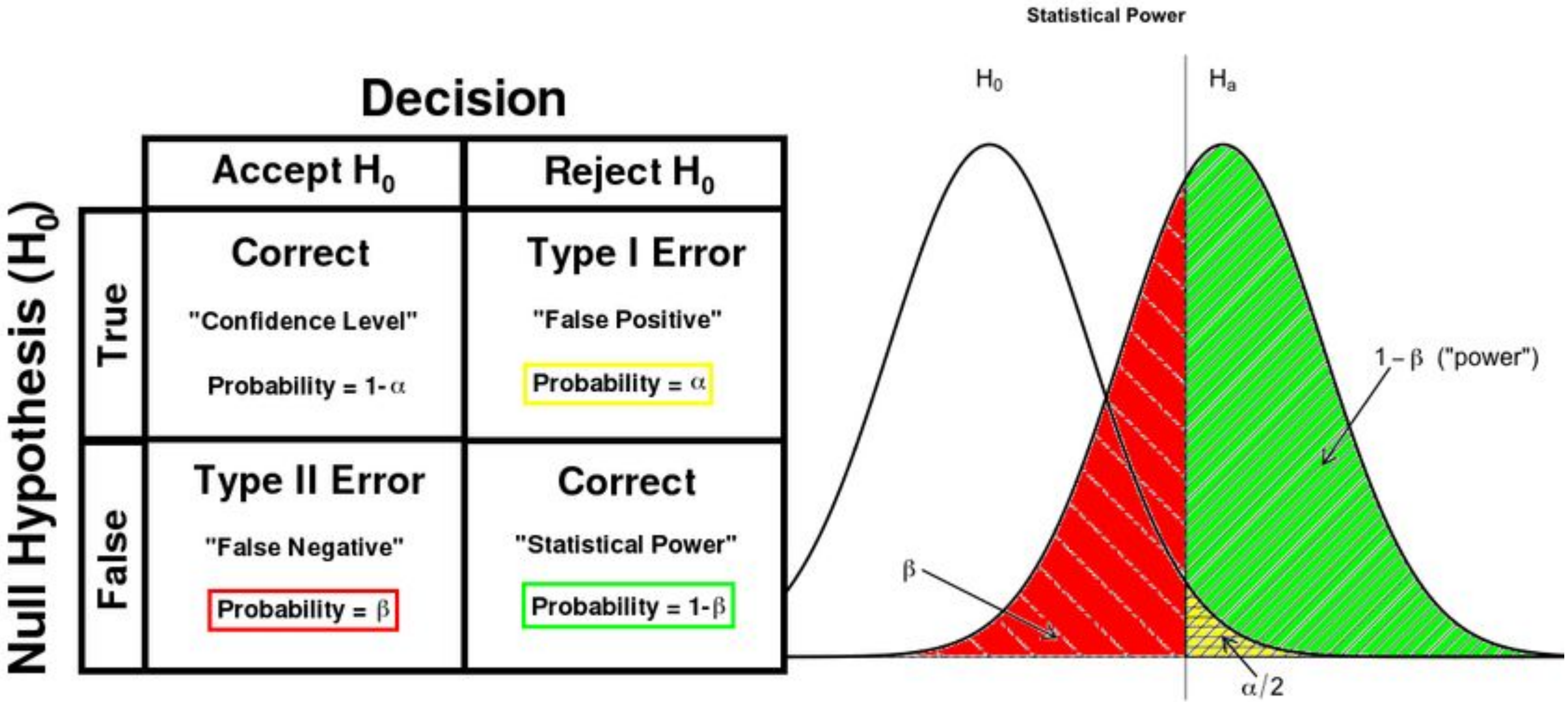
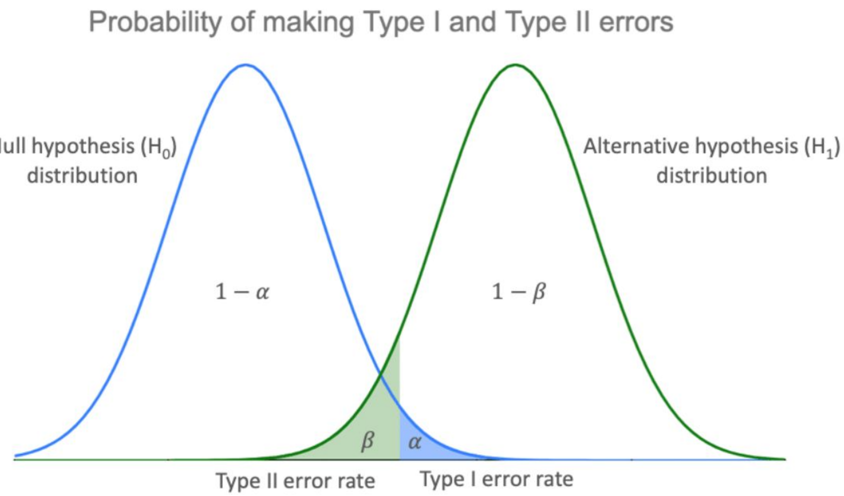


Image Source

Statistical Power

By setting the Type I error rate, you indirectly influence the size of the Type II error rate as well.



It's important to strike a balance between the risks of making Type I and Type II errors. Reducing the alpha always comes at the cost of increasing beta, and vice versa.

Image Source

Statistical Power

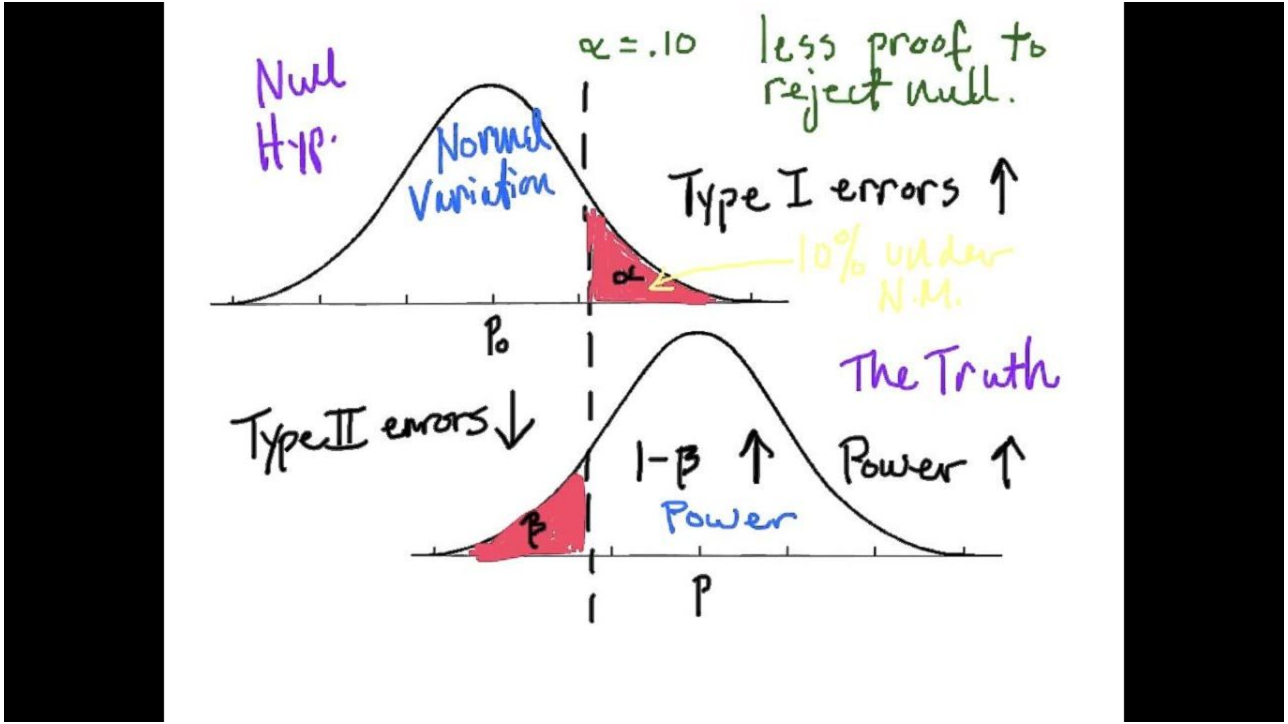


Image Source