Lecture 5:

Hypothesis Testing

Hypothesis testing

- Suppose we want to know is exposure to asbestos is associated with lung disease. We take some rats, randomly divide them into two groups, expose one group to asbestos and leave the second unexposed. Then we compare the disease rate in the two groups. Consider these two hypotheses:

- **The Null Hypothesis:** The disease rate is the same in the two groups. **The Alternative Hypothesis:** The disease rate is not the same in the two groups.

- If the exposed group has much higher rate of disease than the unexposed, we will **reject** the null hypothesis and conclude that evidence favors the alternative hypothesis.

- Sometimes, confidence intervals do the job, with no hyp. test.

Hypothesis testing

- More formally, we divide the parameter space Θ into two **disjoint** sets Θ_0 and Θ_1 and we wish to test:

  $$H_0 : \ \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \ \theta \in \Theta_1$$

  the **null hypothesis** vs the **alternative hypothesis**

- Let X be a r.v. with range $\mathscr{X}$. We test the hypothesis by finding a subset $R \in \mathscr{X}$ called the **rejection region**. If $X \in R$, we reject the null hyp., otherwise – we **retain** (do not reject) the null hyp.

- Usually, the rejection region has form $R = \{x : T(x) > c\}$, where T is a **test statistic** and c is a **critical value**. The problem is to find such a statistic and such a critical value.

Hypothesis testing

- There are two types of errors we can make. **Type I error (false positive)** – rejecting H_0 when it is true. **Type II error (false negative)** – retaining H_0 when H_1 is true.

- **Definition:** The **power function** of a test with rejection region R is defined by

$$\beta(\theta) = \mathbb{P}_\theta(X \in R)$$

The **size** of the test is defined by $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.

A test is said to have **level** α if its size is $\leq \alpha$.

Hypothesis testing

- A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**. A hypothesis of the form $\theta > \theta_0$ (or $\theta < \theta_0$) is called a **composite hypothesis**.

- A test of the form

$$H_0: \ \theta = \theta_0 \quad \text{versus} \quad H_1: \ \theta \neq \theta_0$$

is called a **two-sided test**. A test of the form

$$H_0: \ \theta \leq \theta_0 \quad \text{versus} \quad H_1: \ \theta > \theta_0 \qquad \text{(or vice versa)}$$

is called a **one-sided test**. Two-sided tests are most common.

Hypothesis testing

- **Example**: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$ where σ is known. We want to test $H_0: \ \mu \leq 0$ versus $H_1: \ \mu > 0$. Hence, $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Consider the test:

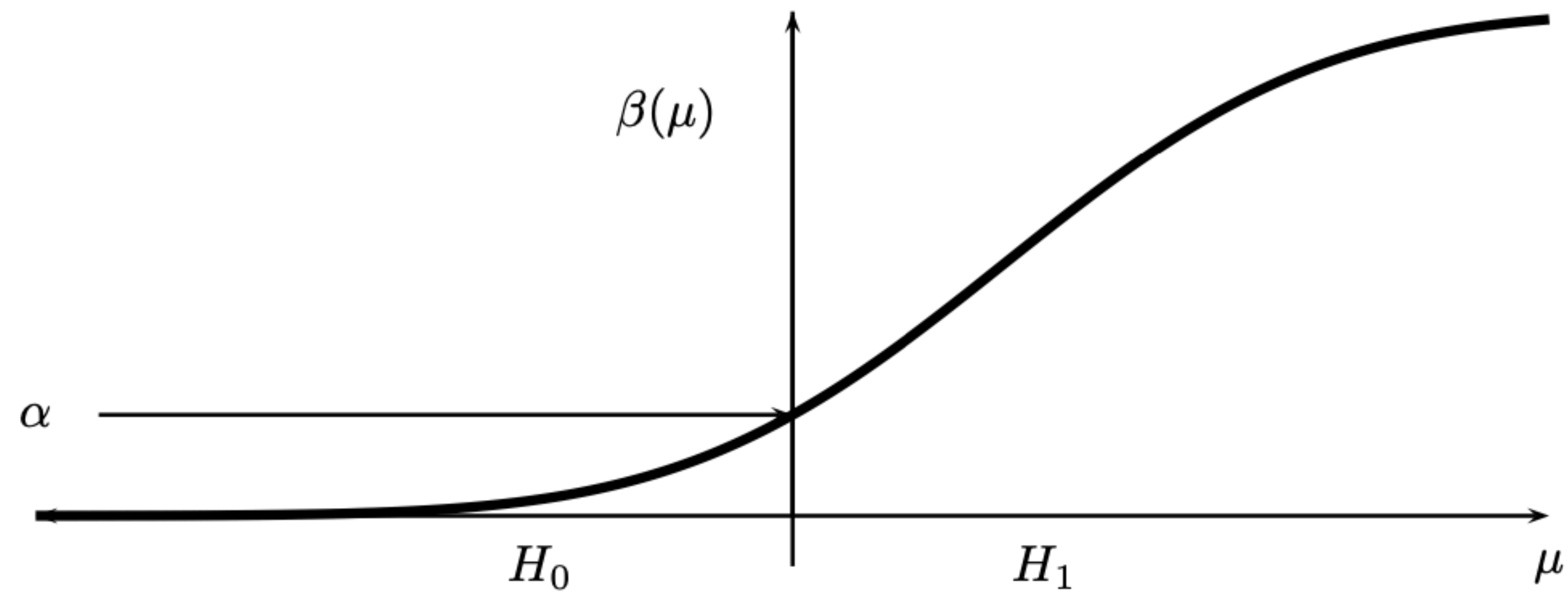
  reject H_0 if $T > c$, where $T = \overline{X}$.

  The rejection region is $R = \left\{(x_1, \ldots, x_n): \ T(x_1, \ldots, x_n) > c\right\}$.

  The power function is

$$\beta(\mu) = \mathbb{P}_\mu(\overline{X} > c) = \mathbb{P}\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right)$$

Hypothesis testing



The power function is increasing in μ, hence

$$\text{size} = \sup_{\mu \le 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right)$$

For a size-α test we fix this equal to α and solve for c to get $c = \sigma\Phi^{-1}(1-\alpha)/\sqrt{n}$. So we reject when $\overline{X} > c$.

Hypothesis testing

- It would be desirable to find the test with highest power under H_1, among all size-α tests. Such a test, if exists, is called **most powerful**. Finding such one is hard, and in many cases most powerful tests don't even exist.

- Instead of going into detail about when such most powerful tests exist, we will consider several widely used tests: 1) the Wald test 2) the χ^2 test, 3) the permutation test, 4) the likelihood ratio test, 5) the t-test

The Wald test

The Wald test

- Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ, and let $\hat{\text{se}}$ be the estimated standard error of $\widehat{\theta}$.

- **Definition**: Consider testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Assume that $\widehat{\theta}$ is asymptotically normal:

$$\frac{\widehat{\theta} - \theta_0}{\hat{\text{se}}} \xrightarrow{d} \mathcal{N}(0,1)$$

The size-α **Wald test** is: reject H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\theta} - \theta_0}{\hat{\text{se}}}$$

The Wald test

- **Theorem**: Asymptotically, the Wald test has size α, that is, $\mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) \to \alpha$, as $n \to \infty$.

- **Remark**: An alternative version of the Wald test statistic is $W = (\widehat{\theta} - \theta_0)/\mathrm{se}_0$, where se_0 is the s.e. computed at $\theta = \theta_0$. Both versions of the test are valid.

- **Theorem**: Suppose the true value of θ is $\theta_\star \neq \theta_0$. The power $\beta(\theta_\star)$ – the proba. of correctly rejecting the null hyp. – is (approx.) $1 - \Phi\left((\theta_0 - \theta_\star)/\hat{\mathrm{se}} + z_{\alpha/2}\right) + \Phi\left((\theta_0 - \theta_\star)/\hat{\mathrm{se}} - z_{\alpha/2}\right)$. Recall that $\hat{\mathrm{se}}$ tends to 0 as sample size increases. So 1) power is large if $\theta_\star$ is far from θ_0, 2) power increases with sample size.

The Wald test

- **Example 1**: (Comparing two predictors). We test two prediction algorithms, $X \sim$ Binomial(m, p_1) and $Y \sim$ Binomial(n, p_2) – being the # of their incorrect predictions on test sets of sizes m and n. The null hyp. is $H_0: \ \delta = p_1 - p_2 = 0$ versus $H_1: \ \delta \neq 0$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$, with estimated s.e. $\hat{\text{se}} = \sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{n}}$. So size-$\alpha$ Wald test is to reject H_0 when $|W| > z_{\alpha/2}$ where $W = \dfrac{\widehat{\delta} - 0}{\hat{\text{se}}}$. The power of this test is largest when p_1 is far from p_2 and when the sample size is large.

The Wald test

- **Example 1**: (Comparing two predictors). What if we tested both algorithms on the same test set? Then two samples are not independent! Denote $X_i =$ Ind(alg1 is correct on i-th test case), and Y_i accordingly, and define $D_i = X_i - Y_i$.

  Now $\delta = \mathbb{E}(D_i) = \mathbb{E}(X_i) - \mathbb{E}(Y_i) = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1)$, and the its plug-in estimate is $\widehat{\delta} = \overline{D}$, and $\hat{\text{se}}(\widehat{\delta}) = S/\sqrt{n}$.

  To test $H_0 : \ \delta = 0$ versus $H_1 : \ \delta \neq 0$ we use $W = \widehat{\delta}/\hat{\text{se}}$ and reject H_0 if $|W| > z_{\alpha/2}$.

  This is called **paired comparison**.

The Wald test

- **Example 2**: (Comparing two means). Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two ind. samples from populations with means μ_1 and μ_2. Null hyp. is $H_0 : \ \delta = \mu_1 - \mu_2 = 0$, versus $H_1 : \ \delta \neq 0$. Plug-in estimate is $\widehat{\delta} = \overline{X} - \overline{Y}$ with estimated s.e:
$\hat{\text{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$ (s_1, s_2 – sample variances).

  Size-α Wald test rejects H_0 when $|W| > z_{\alpha/2}$ where
$W = \frac{\widehat{\delta} - 0}{\hat{\text{se}}}$.

The Wald test

- **Example 3**: (Comparing two medians). Same as previous example, but now plug-in estimates are sample medians, and standard error estimate can be obtained with bootstrap.

- **Theorem**: Size-α Wald test rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where $C = (\widehat{\theta} - \hat{\text{se}}\, z_{\alpha/2},\ \widehat{\theta} + \hat{\text{se}}\, z_{\alpha/2})$ – the **confidence interval.**

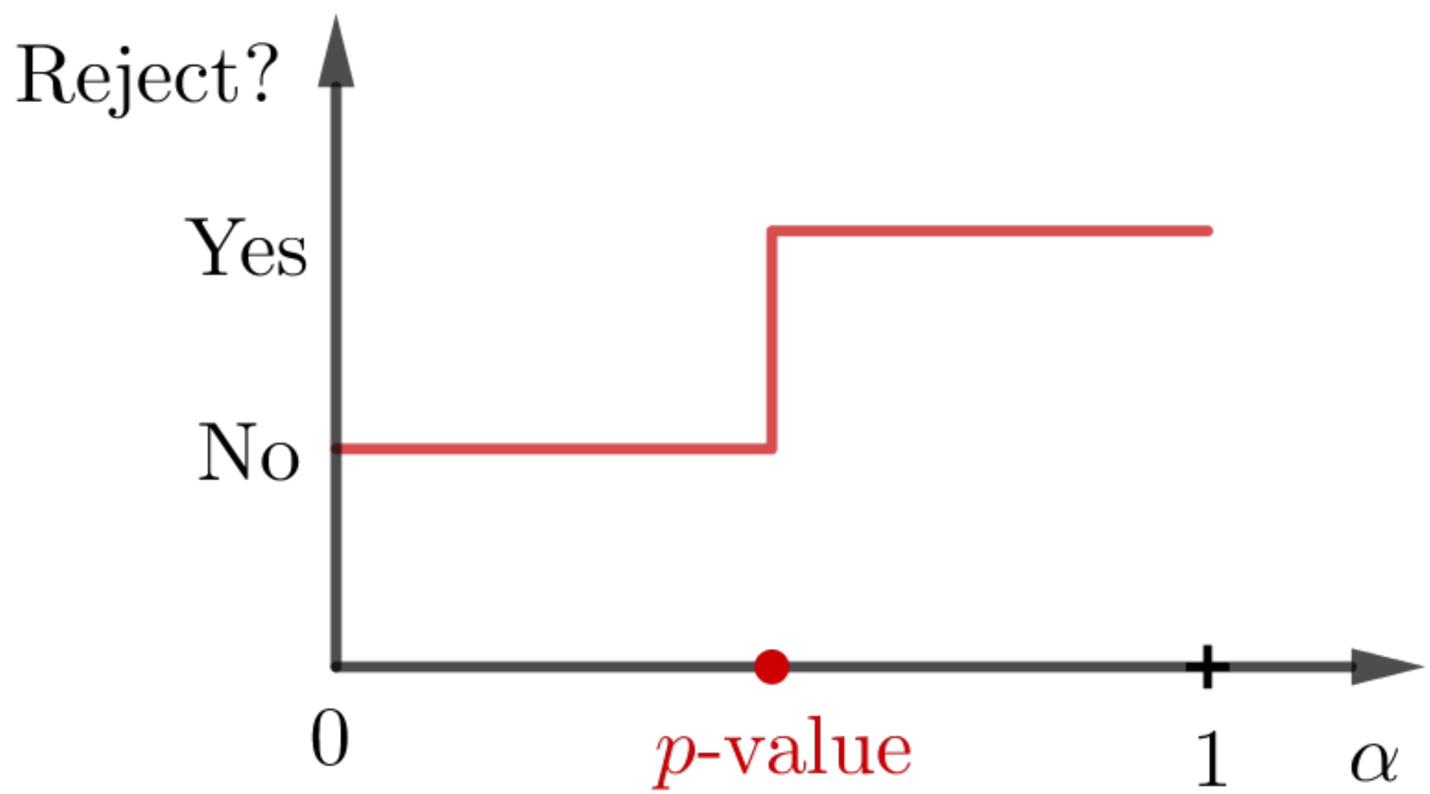
- If we reject H_0 we say the result is **statistically significant**. A result might be statistically significant, but **effect size** (also called scientific significance), $|\theta - \theta_0|$, might be small!

p-values

p-values

- Just reporting "reject H_0" or "retain H_0" is not very informative. Instead, we ask, for every α, whether we reject at that level. The smallest α at which the test rejects is called the **p-value**.

- **Definition**: p-value $= \inf \left\{ \alpha : \; T(X^n) \in R_\alpha \right\}$



p-values

- Informally, p-value is a measure of evidence against H_0: the smaller the p-value, the stronger the evidence against H_0.
- Typically, researchers use the following “evidence scale”:

p-value	evidence
< 0.01	very strong evidence against H_0
0.01-0.05	strong evidence against H_0
0.05-0.10	weak evidence against H_0
> 0.10	little or no evidence against H_0

p-values

- **Warning**: Large p-value is not strong evidence in favor of H_0. It can occur for two reasons: 1) H_0 is true, 2) H_0 is false, but the test has **low power**

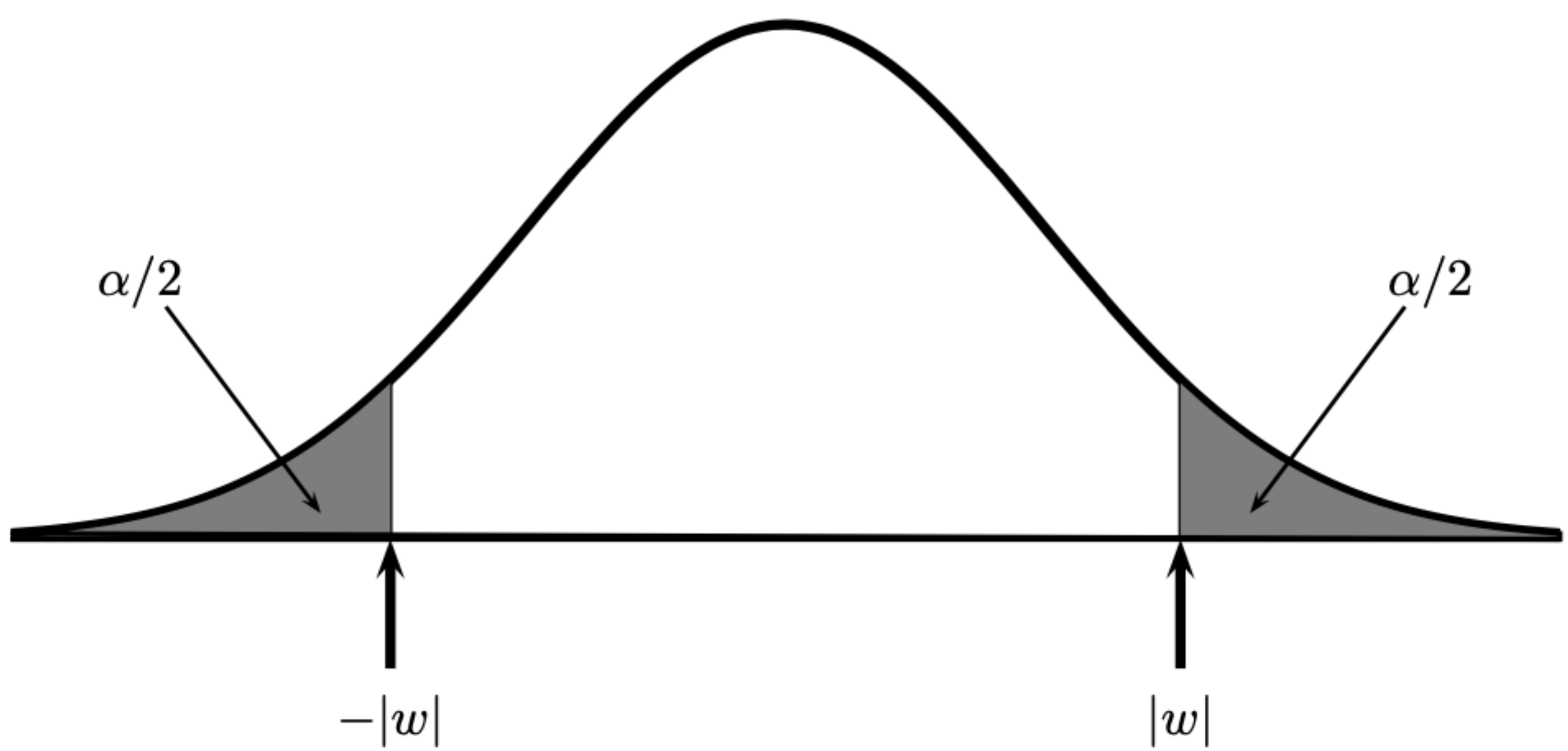
- **Warning**: Do not confuse the p-value with $\mathbb{P}(H_0 \mid \text{data})$!

- Rather than that, the p-value is _the probability (under H_0) of observing a value of the test statistic the same as or more extreme that was actually observed_!

p-values

- **Theorem**: For observed statistic, $w = (\widehat{\theta} - \theta_0)/\hat{\text{se}}$ in Wald test, the p-value is given by

$$\mathbb{P}_{\theta_0}(|W| > |w|) \approx \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|)$$



p-values

- **Theorem**: If the test statistic has a continuous distribution, then under $H_0: \ \theta = \theta_0$, the p-value has a Uniform$(0,1)$ distribution. Therefore, if we reject H_0 when the p-value is less than α, the probability of type I error (false positive) is α.

- In other words, if H_0 is true, the p-value is a random draw from Uniform$(0,1)$. If H_1 is true, the distribution of the p-value will tend to concentrate closer to 0.

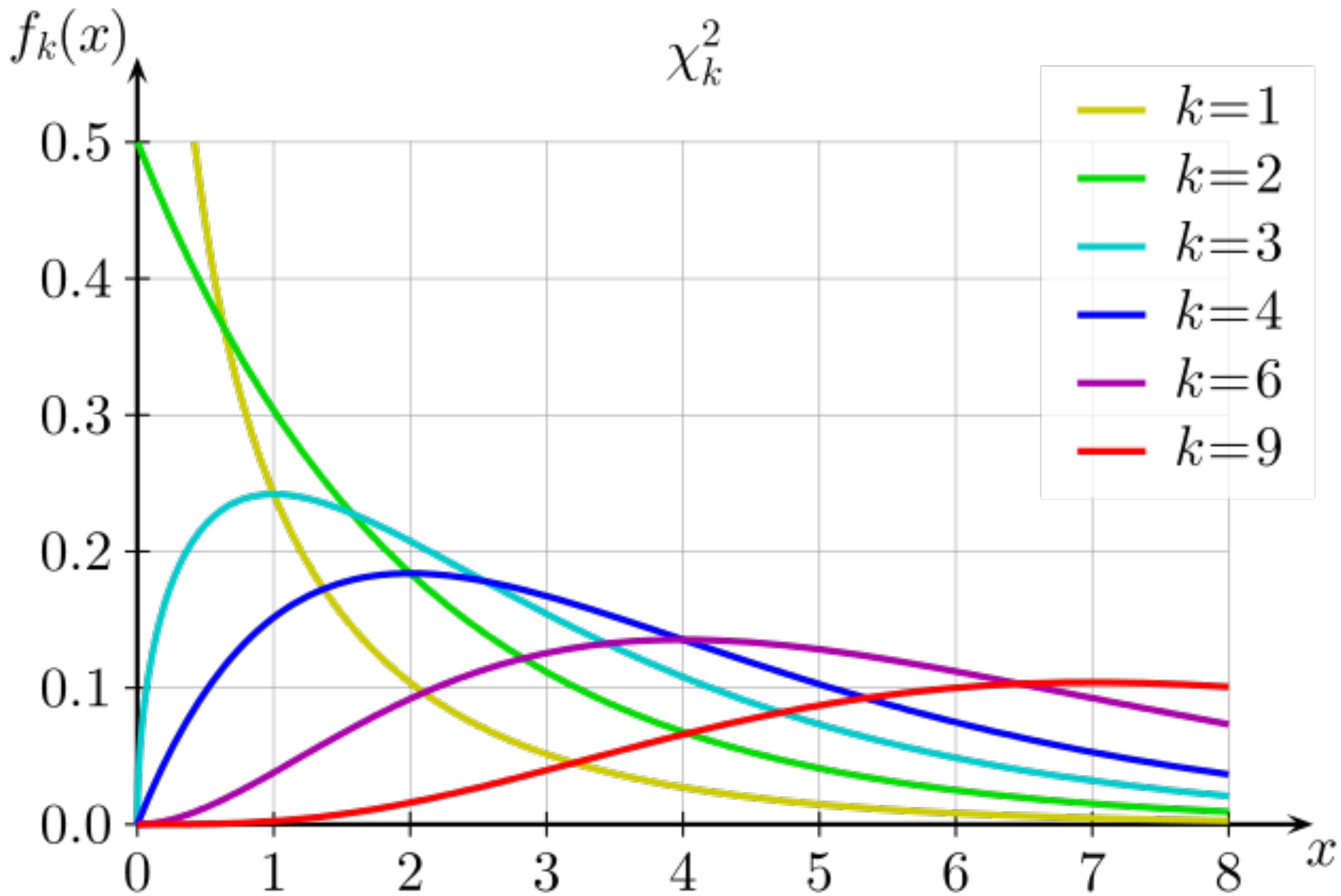
The χ^2 Distribution

The χ^2 distribution

- Let $Z_1, \ldots, Z_k \sim \mathcal{N}(0,1)$. Let $V = \sum_{i=1}^{k} Z_i^2$. Then we say that V has a χ^2-distribution with k **degrees of freedom**, written $V \sim \chi_k^2$
- The pdf is $f(v) = \dfrac{v^{(k/2)-1}\exp(-v/2)}{2^{k/2}\,\Gamma(k/2)}$ for $v > 0$.
- The moments are $\mathbb{E}(V) = k$ and $\mathbb{V}(V) = 2k$.

The χ^2 distribution

- The upper α-quantile is $\chi^2_{k,\alpha} = F^{-1}(1-\alpha)$ where F is the CDF.



The χ^2-test

The χ^2 test

- Pearson's χ^2 test is used for multinomial data. If $X = (X_1, \ldots, X_k)$ has a multinomial (n, p) distribution, then the MLE of p is $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_k) = (X_1/n, \ldots, X_k/n)$.

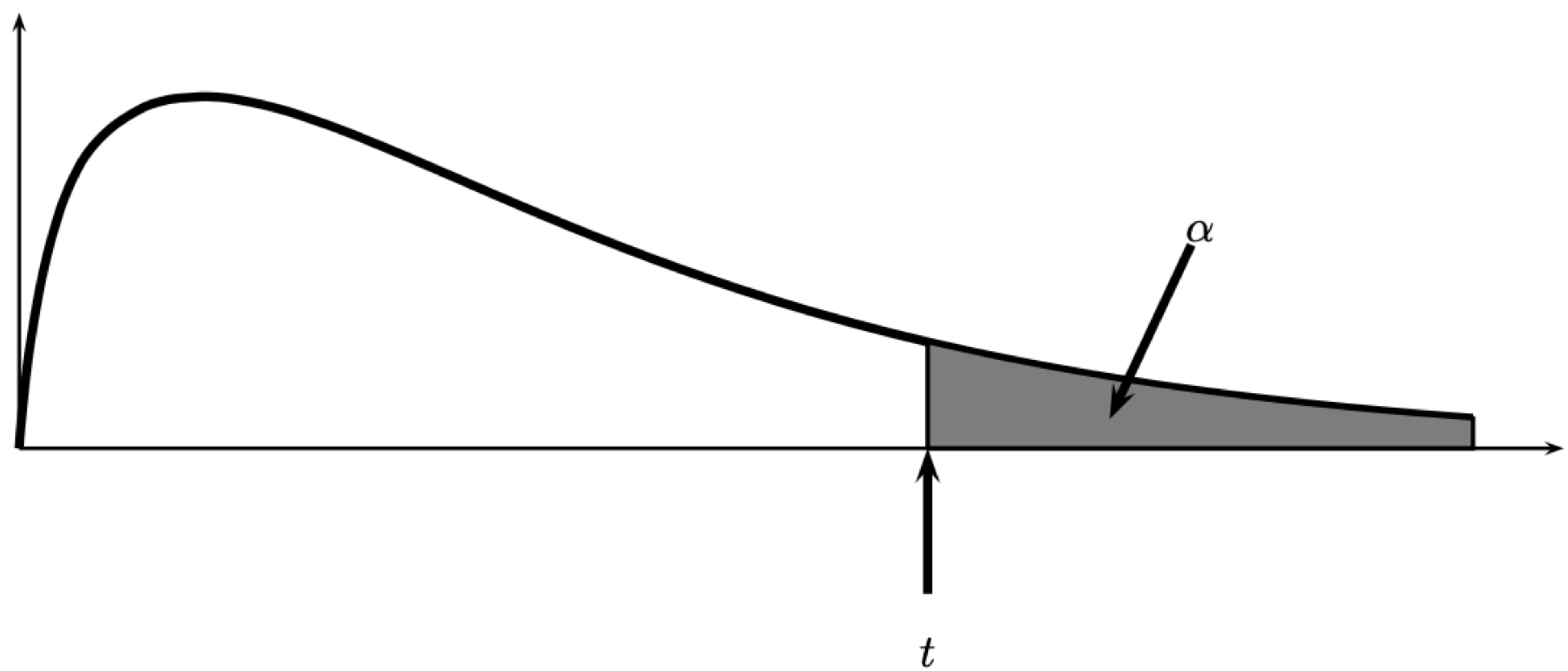
- Let $p_0 = (p_{01}, \ldots, p_{0k})$ be some fixed vector and suppose we want to test $H_0: \ p = p_0$ versus $H_1: \ p \neq p_0$

- **Definition**: Pearson's χ^2 statistic is
$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$ where $E_j = \mathbb{E}(X_j) = np_{0j}$

  is the expectation of X_j under H_0

The χ^2 test

- **Theorem**: Under H_0, we have $T \xrightarrow{d} \chi^2_{k-1}$. Hence the test: reject $T > \chi^2_{k-1,\alpha}$ has asymptotic level α. The p-value is $\mathbb{P}(\chi^2_{k-1} > t)$ where t is the observed value of the test statistic



The χ^2 test

- **Example**: (Mendel's peas). Mendel bred peas with *round yellow* seeds and *wrinkled green* seeds. So there are 4 types of progeny: round/wrinkled, yellow/green. The number of each type is multinomial with $p = (p_1, p_2, p_3, p_4)$. His theory of inheritance predicts that p equals $p_0 \equiv \frac{1}{16}(9,3,3,1)$.

  In $n = 556$ trials he observed $X = (315,101,108,32)$. Since $np_{01} = 312.75$, we have $\chi^2 = \frac{(315 - 312.75)^2}{312.75} + \ldots = 0.47$
  The $\alpha = 0.05$ value for χ^2_3 is 7.815. Since 0.47 < 7.815, we retain the null hyp. The p-value is $\mathbb{P}(\chi^2_3 > 0.47) = 0.93$. So there is not enough evidence to contradict Mendel's theory.

The χ^2 test

- **Remember**: Hypothesis testing is useful to see if there is evidence to reject H_0. It does not prove that H_0 is true! Failure to reject H_0 might occur because 1) H_0 is true, 2) because the test has low power!