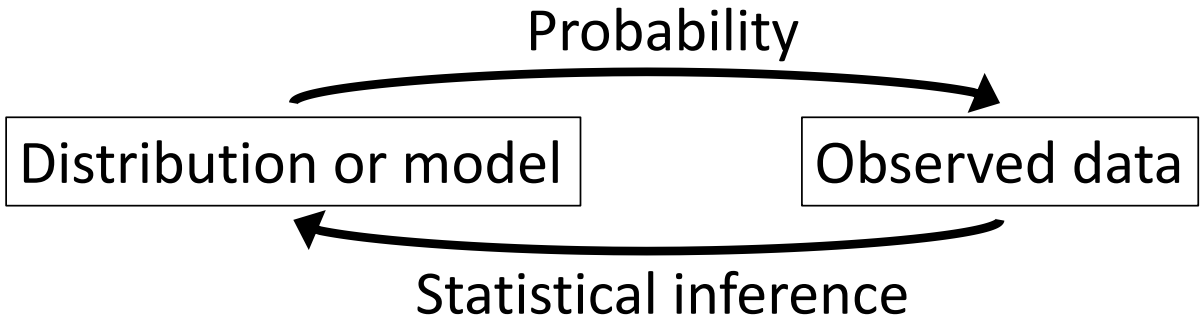
S&DS 242/542: Theory of Statistics

Lecture 5: Null hypotheses, test statistics, and p-values

Testing a simple null hypothesis

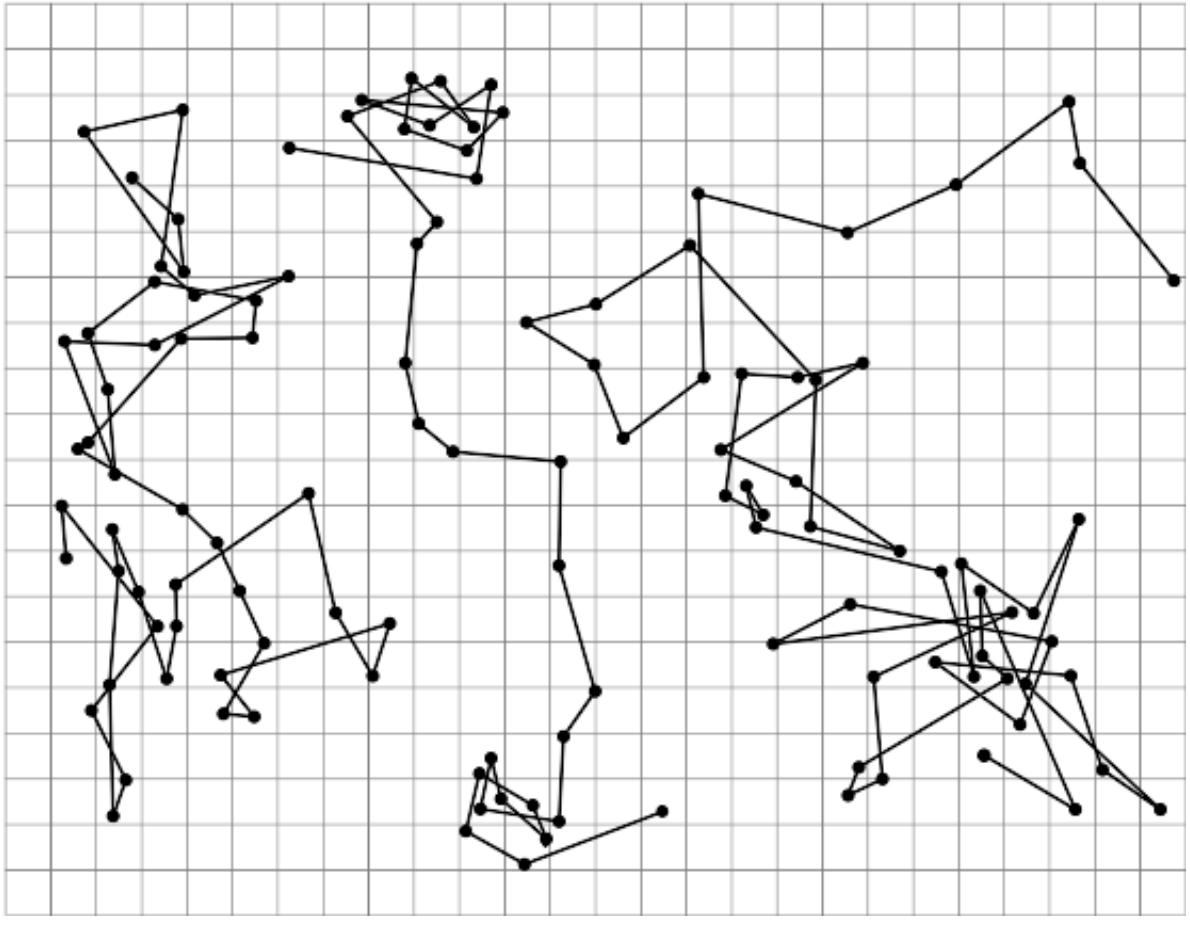


Today: Does my observed data come from a specified distribution?

Example: We roll a 6-sided die n times, and observe $1, 3, 1, 6, 4, 2, 5, 3, \ldots$ Is this a fair die?

Example: Einstein's theory of Brownian motion

Motion of a tiny (radius $\approx 10^{-4}$ cm) particle suspended on the surface of water:



Example: Einstein's theory of Brownian motion

Albert Einstein (1905): Suppose the particle is at position $P_t \in \mathbb{R}^2$ at time t. Then at time $t + \Delta t$, its position $P_{t+\Delta t}$ is random, and has a bivariate normal distribution around P_t. The change in position $P_{t+\Delta t} - P_t$ is independent of the trajectory before time t.

Explanation: In the time period $(t, t + \Delta t)$, the particle is bombarded by water molecules on all sides. Each time a water molecule hits the particle, it moves the particle by a little bit, in a random direction. These collisions are independent and the number of collisions is very large, so their total effect is bivariate normal by the (multivariate) Central Limit Theorem.

In 1905, scientists were debating whether atoms and molecules actually exist. Validation of Einstein's theory was a big step towards proving Dalton's theory of the atom.

Example: Einstein's theory of Brownian motion

More precisely, Einstein predicted:

$$P_{t+\Delta t} - P_t \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2 & 0\\0 & \sigma^2\end{pmatrix}\right)$$

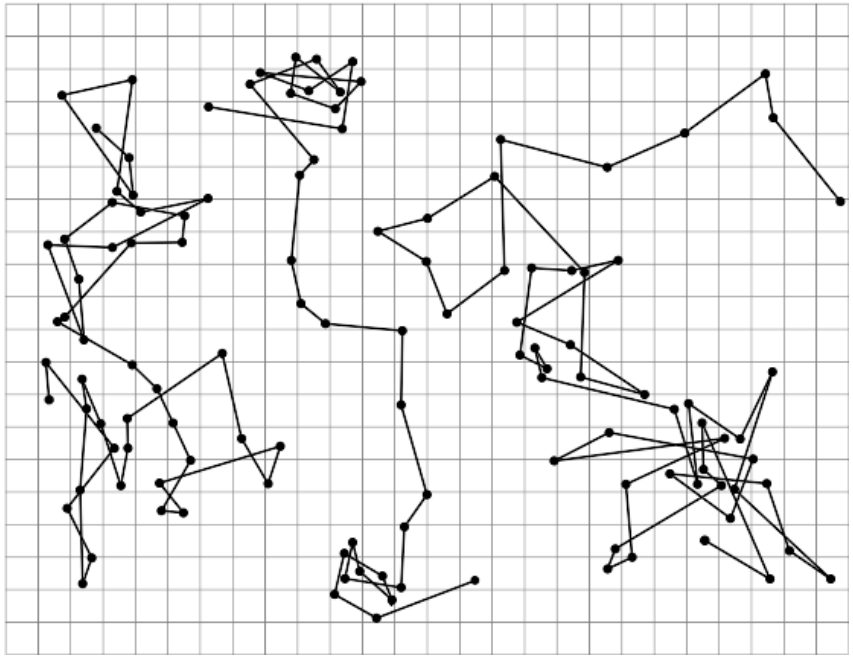
where

$$\sigma^2 = \frac{RT}{3\pi\eta r N_A}(\Delta t).$$

- R: ideal gas constant
- T: absolute temperature
- η: viscosity of water
- r: radius of particle
- N_A: Avogadro's number

Example: Einstein's theory of Brownian motion

Jean Perrin (1909): Measured the position of the particle every 30 seconds, to test Einstein's theory. For his experiment, the variance predicted by Einstein's theory was $\sigma^2 = 2.23 \times 10^{-7}$ cm^2.



Does Perrin's data support Einstein's theory of Brownian motion?

Null and alternative hypotheses

A **hypothesis test** is a binary question about the distribution of the data.

Our goal is to either accept a **null hypothesis** H_0 about this distribution, or reject it in favor of an **alternative hypothesis** H_1.

Today we'll focus on the null hypothesis H_0. We'll think more about the alternative hypothesis H_1 in later lectures.

Null and alternative hypotheses

Example: Let $(X_1, \ldots, X_6)$ be the numbers of 1's through 6's in n rolls of a six-sided die. The hypothesis that the die is fair is the null hypothesis

$$H_0 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, \left(\tfrac{1}{6}, \ldots, \tfrac{1}{6}\right)\right).$$

We might wish to test this null hypothesis against the alternative hypothesis that the die is not fair,

$$H_1 : (X_1, \ldots, X_6) \sim \text{Multinomial}(n, (p_1, \ldots, p_6))$$
$$\text{for some probability vector } (p_1, \ldots, p_6) \neq \left(\tfrac{1}{6}, \ldots, \tfrac{1}{6}\right).$$

Null and alternative hypotheses

Example: Let

$$\begin{aligned}(X_1, Y_1) &= P_1 - P_0,\\ (X_2, Y_2) &= P_2 - P_1,\\ &\vdots\\ (X_n, Y_n) &= P_n - P_{n-1}\end{aligned}$$

be the displacements measured in Perrin's experiment, where $P_0, P_1, P_2 \ldots \in \mathbb{R}^2$ are the positions (in cm) every 30 seconds.

Einstein's theory is the null hypothesis

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2.23\text{e}{-7} & 0\\ 0 & 2.23\text{e}{-7}\end{pmatrix}\right)$$

Null and alternative hypotheses

There might be various alternative hypotheses that we are interested in testing this against.

The theory is qualitatively correct, but the variance is wrong:

$$H_1 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

for some $\sigma^2 \neq 2.23\mathrm{e}{-7}$.

Or maybe the drift is not normal:

$H_1 : (X_1, Y_1), \ldots, (X_n, Y_n)$ are IID from a distribution that is not bivariate normal

Or maybe these displacements are not independent, etc.

Which is the null and which is the alternative?

We will discuss the classical *Neyman-Pearson paradigm* for hypothesis testing. In this approach, the null and alternative hypotheses are not treated symmetrically.

The question we will ask is: Does the data provide sufficiently strong evidence to reject H_0, in favor of H_1?

This means that the "default" assumption is that H_0 is true. The burden is on the investigator to convincingly demonstrate that H_0 is *false*, not that it is *true*.

Which hypothesis should be the null?

Example: In clinical trials for drugs, typically

$$H_0 : \text{Drug has no treatment effect}$$
$$H_1 : \text{Drug has a treatment effect}$$

The burden is on the investigator to demonstrate, using data from the clinical trial, that the drug is effective.

Example: Does ESP (extrasensory perception) exist? Most studies of this were conducted with

$$H_0 : \text{ESP does not exist}$$
$$H_1 : \text{ESP exists}$$

The burden is on the investigator to show that ESP exists.

Test statistics

A **test statistic** T is any statistic computed from the data, such that an extreme value for T (too large, or too small, or either too large or too small) provides evidence against H_0, in favor of H_1.

Example: Let $(X_1, \ldots, X_6)$ count the results from n rolls of a six-sided die. One possible test statistic is

$$T = \left(\frac{X_1}{n} - \frac{1}{6}\right)^2 + \ldots + \left(\frac{X_6}{n} - \frac{1}{6}\right)^2.$$

Large values of T provide evidence against the null hypothesis that the die is fair,

$$H_0 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, \left(\tfrac{1}{6}, \ldots, \tfrac{1}{6}\right)\right).$$

Test statistics

Let's conceptually separate the problem of hypothesis testing into two questions:

1. **How can we design good test statistics T?**
2. How can we decide if H_0 is true or false based on T?

Test statistics

Example: Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the displacements from Perrin's experiment. For testing

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2.23\text{e}{-7} & 0 \\ 0 & 2.23\text{e}{-7} \end{pmatrix} \right)$$

against the alternative $\sigma^2 \neq 2.23\text{e}{-7}$, one idea for a test statistic is

$$\bar{R} = \frac{1}{n}\Big(\underbrace{X_1^2 + Y_1^2}_{=R_1} + \underbrace{X_2^2 + Y_2^2}_{=R_2} + \ldots + \underbrace{X_n^2 + Y_n^2}_{=R_n} \Big)$$

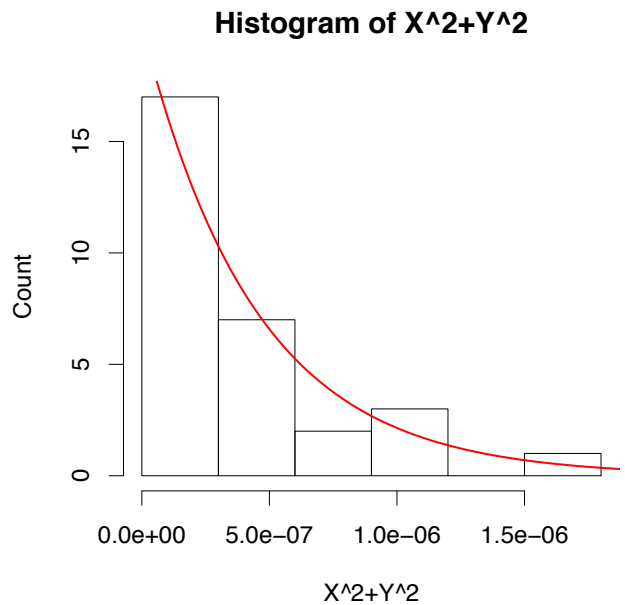
Under Einstein's theory, we should have

$$\mathbb{E}[\bar{R}] = \mathbb{E}[X_i^2 + Y_i^2] = 4.46\text{e}{-7}.$$

Values of $\bar{R}$ much larger or smaller than $4.46\text{e}{-7}$ may indicate that the predicted variance $\sigma^2 = 2.23\text{e}{-7}$ is wrong. However, this statistic $\bar{R}$ may not be able to detect departures from normality.

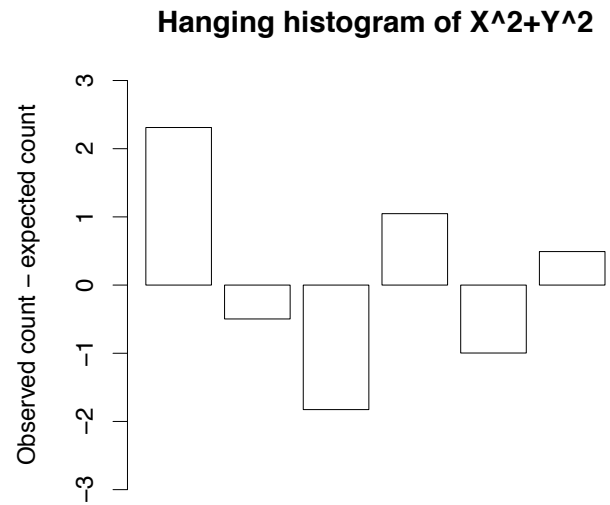
Test statistics from histograms

Let's consider again the value $R_i = X_i^2 + Y_i^2$. If Einstein's theory were correct, then this should be the sum-of-squares of two IID normals, which has the distribution $2.23\mathrm{e}{-7} \cdot \chi^2_2$. Instead of just looking at the mean $\bar{R}$, we can plot a histogram of the values $R_1, \ldots, R_n$ to assess goodness-of-fit to the χ^2-distribution.



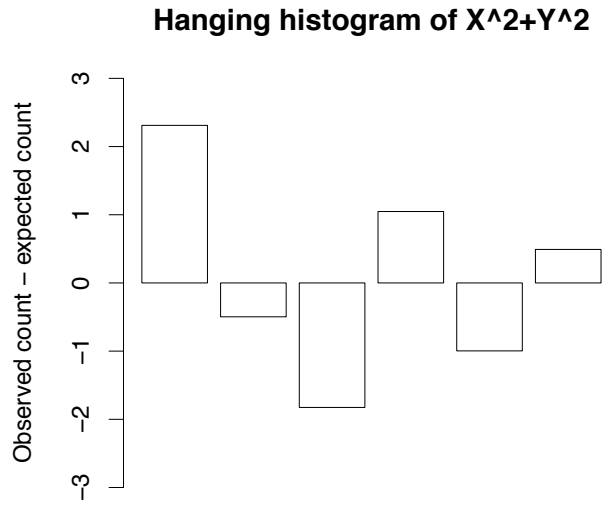
Test statistics from histograms

Deviations from this distribution are better visualized by a **hanging histogram**. This plots $O_i - E_i$ for each histogram bin, where O_i is the observed count for bin i and E_i is the theoretical expected count under the $2.23\mathrm{e}{-7} \cdot \chi^2_2$ distribution:



We may compute from this a test statistic $T = \sum_i (O_i - E_i)^2$. If T is too large, then this may indicate that $R_1, \ldots, R_n$ do not have the distribution $2.23\mathrm{e}{-7} \cdot \chi^2_2$, so Einstein's theory may be wrong.

Test statistics from histograms



The bars of this hanging histogram are larger on the left and smaller on the right. Should the bars on the left provide more evidence against the hypothesized $2.23\mathrm{e}{-7} \cdot \chi^2_2$ distribution?

Test statistics from histograms

Not necessarily: Let p_i be total probability of bin i. If $f(x)$ is the PDF of $2.23\mathrm{e}{-7} \cdot \chi^2_2$, then

$$p_i = \int_{\text{bin } i} f(x)dx.$$

The observed count for bin i is $O_i \sim \text{Binomial}(n, p_i)$, and the expected count is $E_i = np_i = \mathbb{E}[O_i]$. So

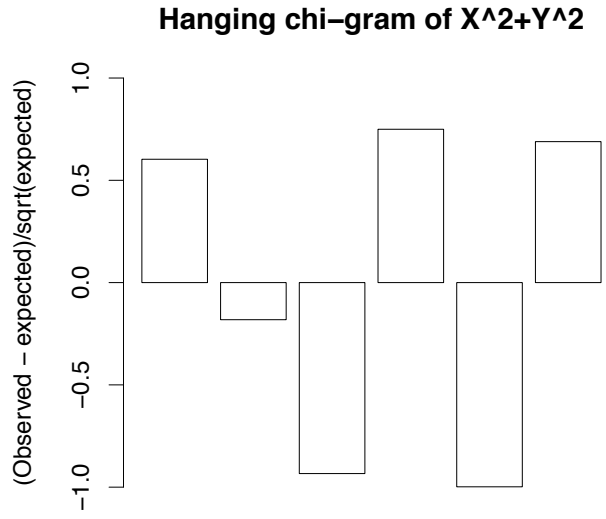
$$\mathbb{E}[(O_i - E_i)^2] = \text{Var}[O_i] = np_i(1 - p_i).$$

Thus $(O_i - E_i)$ is more variable if the bin probability p_i is close to $1/2$, and less variable if it is close to 0 or 1.

If p_i for each individual bin is small, then $\text{Var}[O_i] \approx np_i = E_i$, which is smaller for bins with smaller expected counts E_i.

Test statistics from histograms

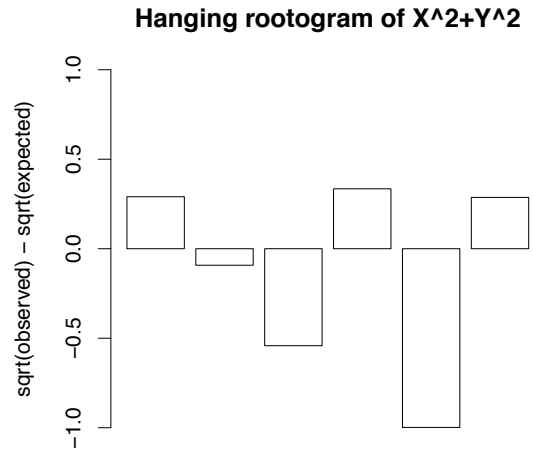
To balance contributions from low- and high-probability bins, we can "stabilize the variance" by plotting $\frac{O_i - E_i}{\sqrt{E_i}}$, so that $\mathbb{E}[(\frac{O_i - E_i}{\sqrt{E_i}})^2] \approx 1$ for all bins. This is called a **hanging chi-gram**.



The reweighted test statistic $T = \sum_i \frac{(O_i - E_i)^2}{E_i}$ is called **Pearson's chi-squared statistic for goodness of fit**.

Test statistics from histograms

Alternatively, to stabilize the variance, we can plot $\sqrt{O_i} - \sqrt{E_i}$. This is called **Tukey's hanging rootogram**.

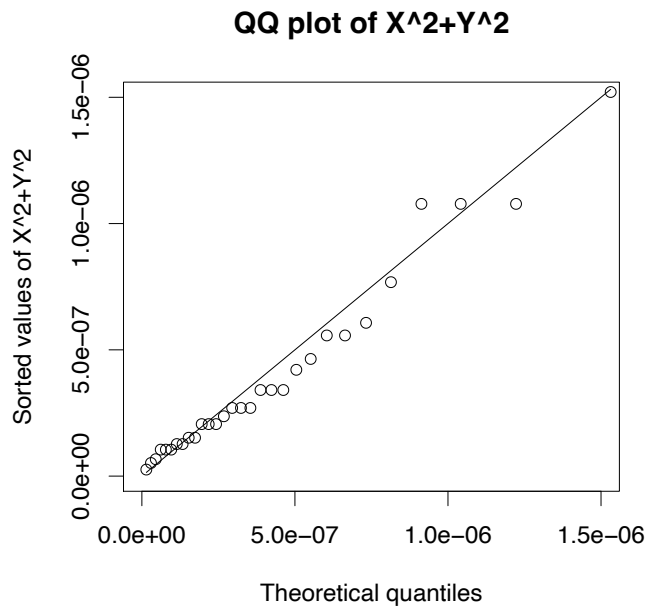


Taylor expansion of $\sqrt{x}$ around $x = E_i$ yields $\sqrt{O_i} - \sqrt{E_i} \approx \frac{O_i - E_i}{2\sqrt{E_i}}$ so this is similar to the hanging chi-gram when $O_i - E_i$ are small.

This motivates another test statistic $T = \sum_i (\sqrt{O_i} - \sqrt{E_i})^2$.

Test statistics from QQ plots

A different visualization of goodness-of-fit is the **QQ plot** (quantile-quantile plot). This plots the sorted values of $R_1, R_2, \ldots, R_n$ against the $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$ quantiles* of their hypothesized distribution, $2.23\text{e}{-7} \cdot \chi^2_2$ in our case:



Values far from the diagonal line $y = x$ provide evidence against the hypothesized distribution.

*It is common to use instead $\frac{1}{n+1}, \frac{2}{n+1} \ldots, \frac{n}{n+1}$ to avoid a largest quantile of ∞.

Test statistics from QQ plots

How can we get a test statistic from a QQ plot? One idea is to take the maximum vertical deviation from the $y = x$ line.

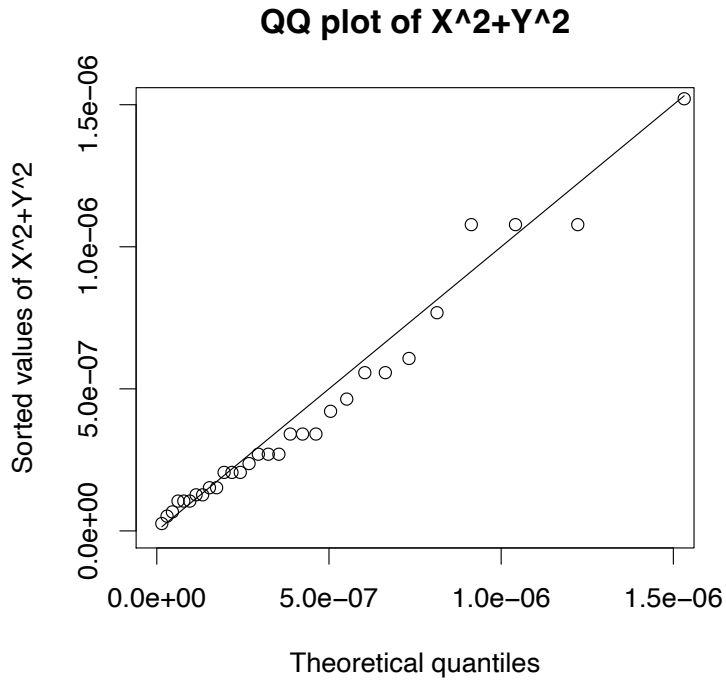
Let $R_{(1)} < \ldots < R_{(n)}$ be the sorted values of $R_1, \ldots, R_n$. Then this maximum vertical deviation is

$$T = \max_{i=1}^{n} \left| R_{(i)} - F^{-1}\left(\frac{i}{n}\right) \right|$$

Here, F^{-1} is the quantile function of the hypothesized distribution (the inverse function of its CDF), so $F^{-1}(\frac{i}{n})$ is its $\frac{i}{n}$-quantile.

[Instead of taking the maximum deviation, one may also take the average deviation, the sum-of-squared-deviations, etc.]

Test statistics from QQ plots



The deviations are larger to the right of the QQ plot, where the observed values are more spaced out. Should these points provide more evidence against the $2.23\mathrm{e}{-7} \cdot \chi^2_2$ distribution?

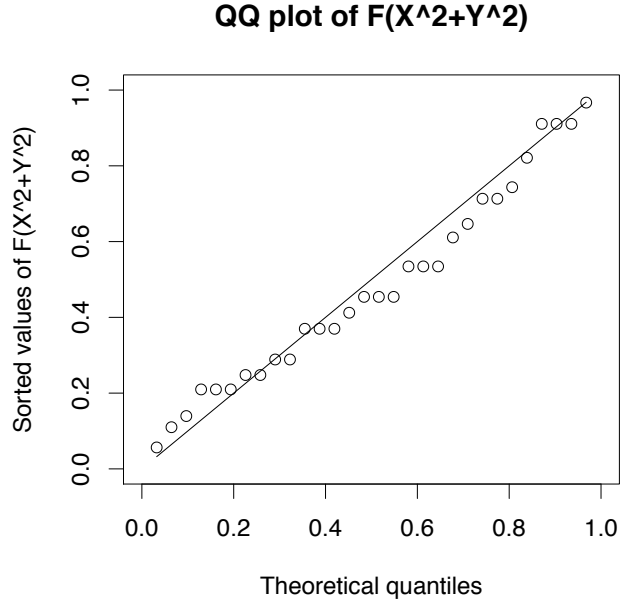
Not necessarily. The sorted values of R_i are closer together on the left, so they are less variable.

Test statistics from QQ plots

We may stabilize the spacings between quantiles by considering instead

$$T = \max_{i=1}^{n} \left| F(R_{(i)}) - \frac{i}{n} \right|.$$

This is the maximum vertical deviation of a QQ-plot of the sorted values of $F(R_1), F(R_2), \ldots, F(R_n)$ against the values $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$:



This is the **Kolmogorov-Smirnov statistic for goodness of fit**.

Conducting the hypothesis test

The choice of test statistic depends on the null hypothesis we wish to test, how we wish to summarize the evidence in our data, and the alternative hypotheses we wish to test against.

We'll see a few principles for choosing test statistics and some more examples in later lectures.

1. How can we design good test statistics T?
2. **How can we decide if H_0 is true or false based on T?**

The null distribution

We usually cannot assert with 100% confidence that H_0 is false. But we can compute T from our data, and compare this value with the sampling distribution of T if H_0 were true. This is called the **null distribution** of T.

Example: Let $X_1, \ldots, X_n$ be IID normal variables. We wish to test

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, 1)$$

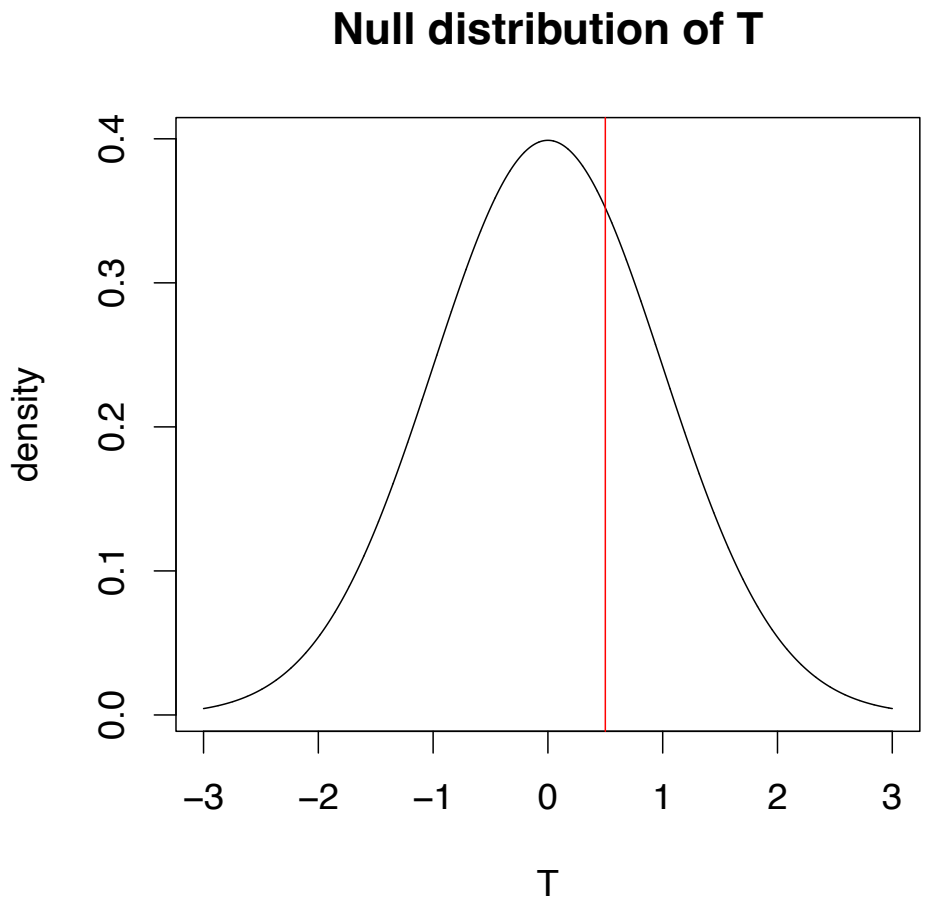
$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\theta, 1) \text{ for some } \theta > 0$$

Consider the test statistic $T = \sqrt{n}\,\bar{X}$. If H_0 were true, then

$$T \sim \mathcal{N}(0, 1)$$

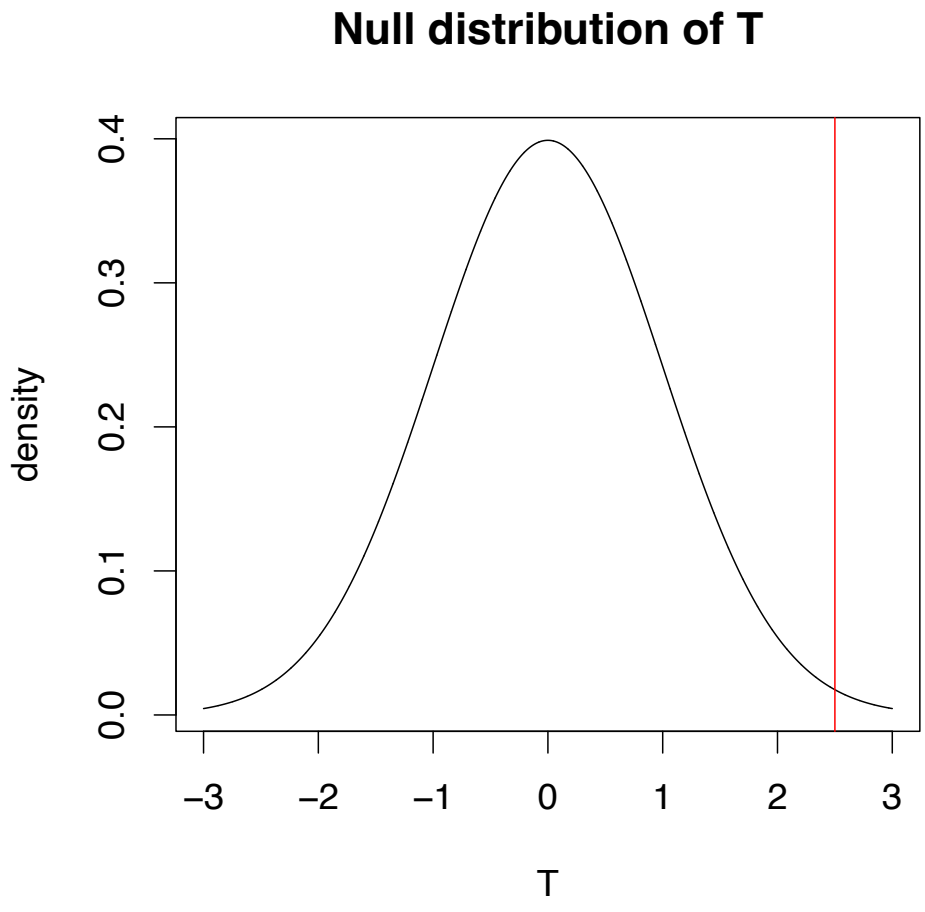
This is the null distribution of T.

The null distribution



If we observe $T = 0.5$, this would not provide strong evidence against H_0 in favor of H_1. In this case we might accept H_0.

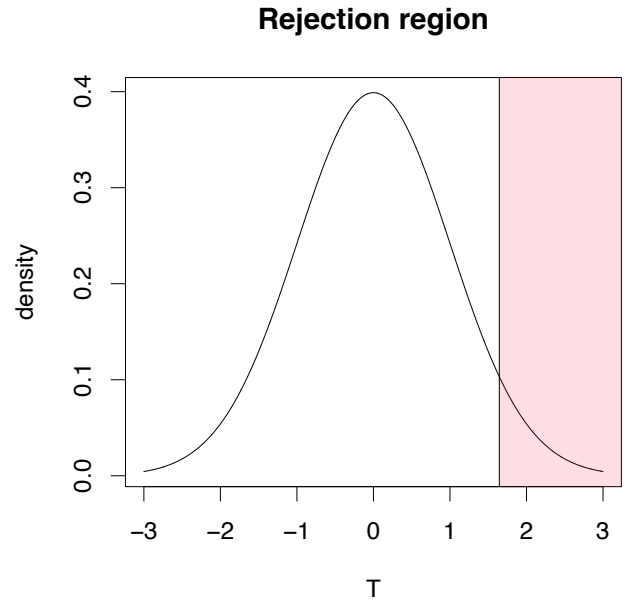
The null distribution



If we observe $T = 2.5$, this would provide stronger evidence against H_0. In this case we might reject H_0 in favor of H_1.

Rejection and acceptance regions

For a test statistic T, we divide its possible values into a **rejection region** and an **acceptance region**, for rejecting/accepting H_0.



In the previous example, large values of T provide evidence against H_0 in favor of H_1, so we might define our rejection region as all values greater than some threshold, as depicted in red.

Significance level and Type I error

The probability of **Type I error** is the probability that we wrongly reject H_0, when H_0 is in fact true:

$$\mathbb{P}[\text{Type I error}] = \mathbb{P}_{H_0}[\text{reject } H_0]$$

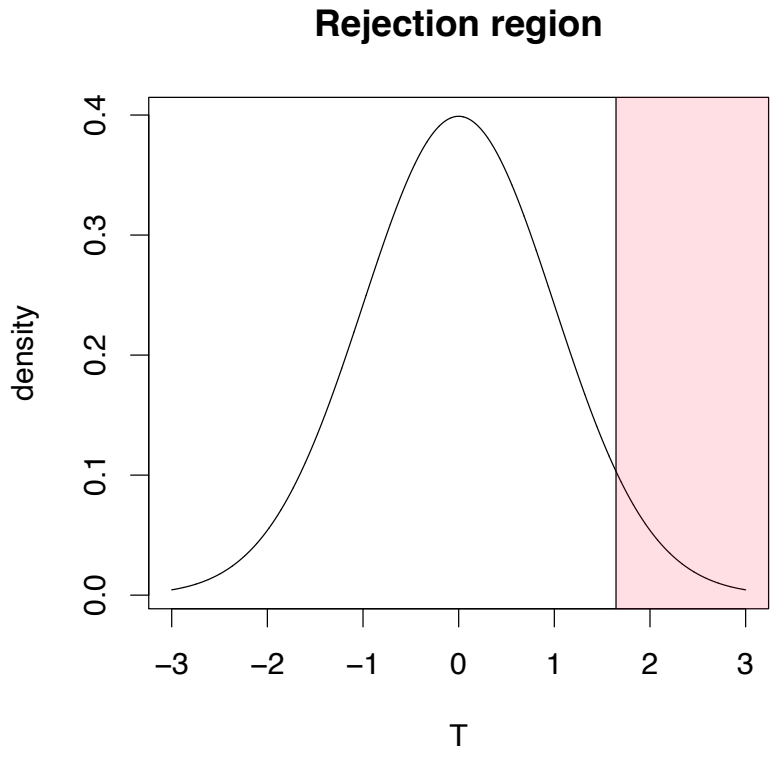
We write $\mathbb{P}_{H_0}$ to mean that this probability is computed under the assumption that H_0 is true, i.e. using the null distribution of T.

We may choose the rejection region for T to ensure that

$$\mathbb{P}[\text{Type I error}] \leq \alpha$$

for a specified **significance level** $\alpha \in (0, 1)$ of the test.

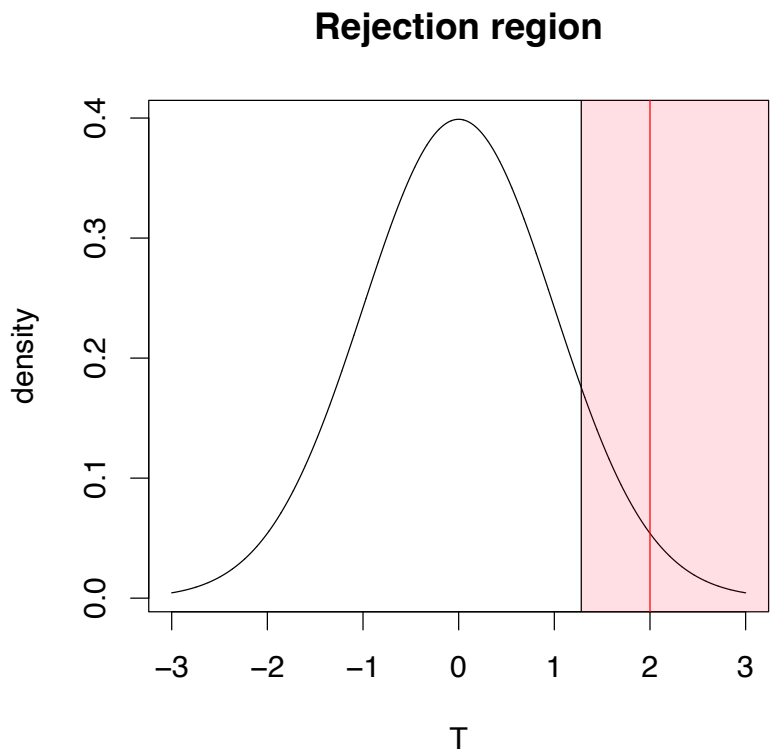
Signficance level and Type I error



Here, the null distribution is $\mathcal{N}(0,1)$. If we wish to perform the test at significance level α, we may set the threshold of the rejection region to be the "upper-α point" (i.e. the $(1-\alpha)^{\text{th}}$ quantile) of the $\mathcal{N}(0,1)$ distribution.

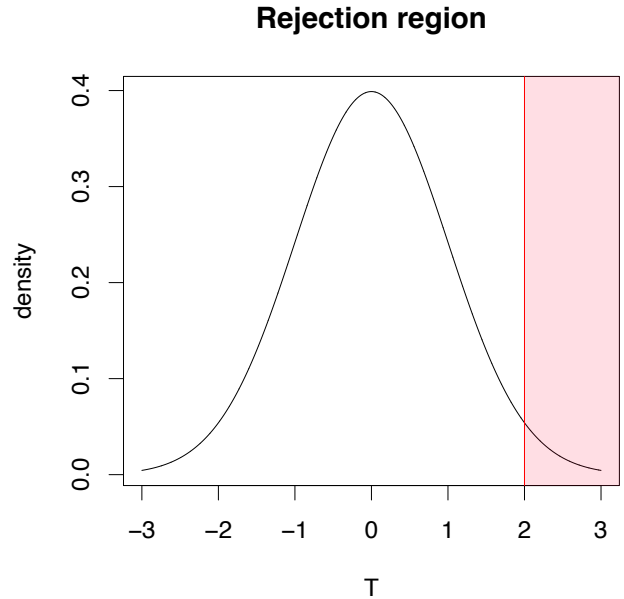
p-values

Oftentimes, we do not want to fix a specific significance level. We may instead ask the question: Would the test reject H_0 at level $\alpha = 0.01$? At level $\alpha = 0.05$? At level $\alpha = 0.1$? ...



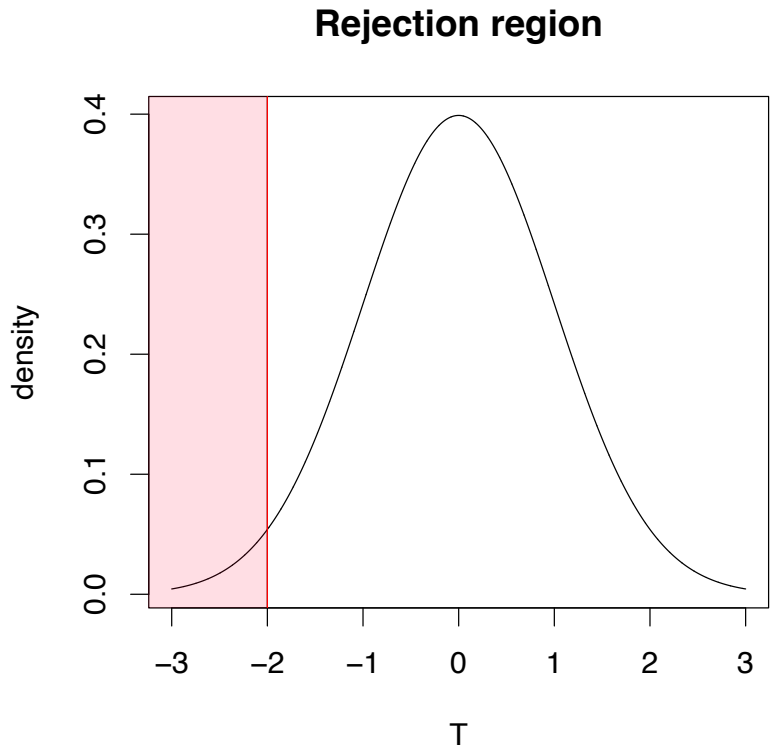
p-values

The smallest signifance level at which we *would* reject H_0 is called the **p-value** for our test. This provides a quantitative measure of the extent to which the data supports the null hypothesis.



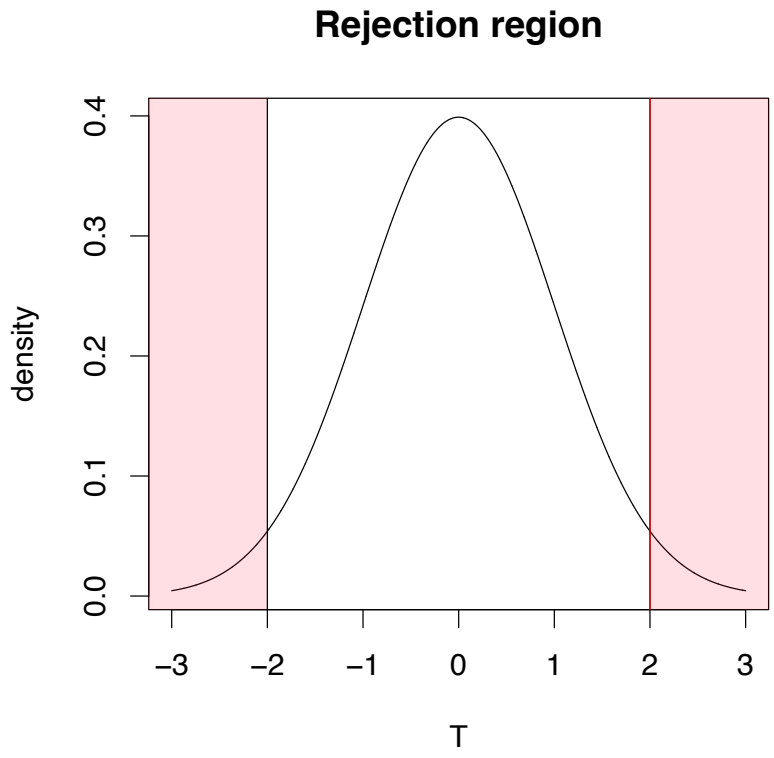
In this example, the p-value is $\mathbb{P}_{H_0}[T \geq t_{\text{obs}}]$, where t_{obs} is our observed value of the test statistic T. In other words, it is the total probability that the null distribution assigns to values $\geq t_{\text{obs}}$.

p-values



If our alternative is $H_1 : \theta < 0$ and our test rejects H_0 for small values of T, then the p-value is $\mathbb{P}_{H_0}[T \leq t_{\text{obs}}]$, the total probability that the null distribution assigns to values $\leq t_{\text{obs}}$.

p-values



If our alternative is $H_1 : \theta \neq 0$ and we perform a "two-sided" test that rejects H_0 for large values of $|T|$, then the p-value is $\mathbb{P}_{H_0}[|T| \geq |t_{\text{obs}}|]$, the total probability that the null distribution assigns to both values $\geq |t_{\text{obs}}|$ and $\leq -|t_{\text{obs}}|$.

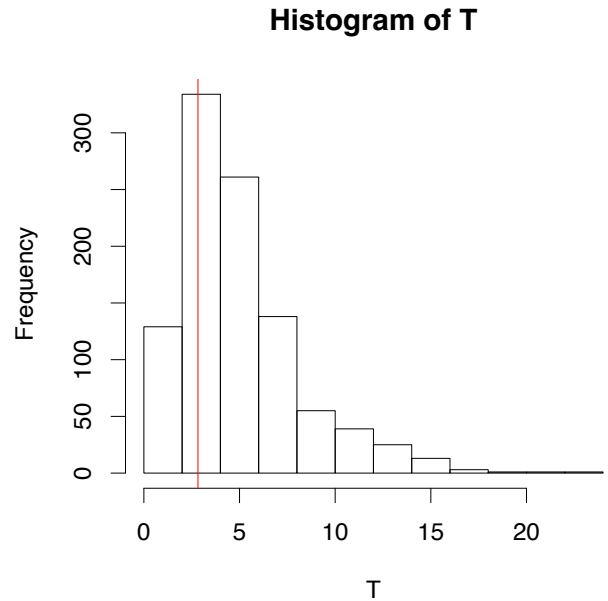
Determining the null distribution

To determine the rejection region for a test statistic T at a given significance level, or to compute its p-value, we must know its null distribution — what typical values of T look like if H_0 were true.

- Sometimes we can derive the null distribution exactly. In the previous example, the null distribution was $\mathcal{N}(0, 1)$.
- Sometimes we can derive a large-sample approximation, using the LLN, CLT, and tools that we discussed last lecture.
- Sometimes we can approximate the null distribution by simulation.

Simulated null distribution for Perrin's data

Let T be the previously discussed Pearson's chi-squared statistic for goodness-of-fit (computed from 6 histogram bins), on the values $R_1, \ldots, R_n$ from Perrin's experiments.



This is the null distribution of T obtained from 1000 simulations of particle paths that follow Einstein's theory of Brownian motion. The observed value $t_{\text{obs}} = 2.83$ for Perrin's actual data is in red.

p-value for Perrin's data

The p-value is the right tail probability $P = \mathbb{P}_{H_0}[T \geq 2.83]$.

In our simulations, 75.4% of simulated values for T were larger than $t_{\text{obs}} = 2.83$, so we may approximate $P \approx 0.754$. This indicates little evidence against Einstein's theory.

[In this example of Pearson's chi-squared test, there is in fact a large-sample χ^2-approximation to the null distribution of T. This is usually used instead of simulation to assess statistical significance and compute a p-value.]