STATS 200: Introduction to Statistical Inference

Lecture 5: Testing a simple null hypothesis

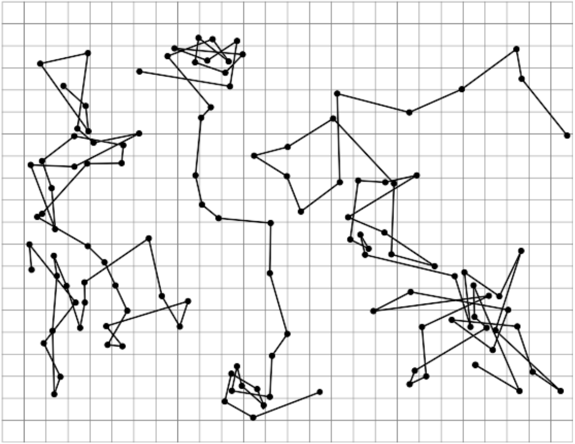
Statistical inference = Probability^{-1}

Today: Does my data come from a prescribed distribution, F?
This is oftentimes called testing **goodness of fit**.

Example: You 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 in water:



Example: Einstein's theory of Brownian motion

Albert Einstein (1905): $P_{t+\Delta t} \sim \mathcal{N}\left(P_t, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$, where

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

- P_t: position of particle at time 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

Albert Einstein (1905): $P_{t+\Delta t} \sim \mathcal{N}\left(P_t, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$, where

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

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

Jean Perrin (1909): Measured the position of a particle every 30 seconds to verify Einstein's theory (and to compute N_A). For his experiment, $\sigma^2 = 2.23 \times 10^{-7}$ cm^2.

Does Perrin's data fit with Einstein's model?

Null and alternative hypotheses

A **hypothesis test** is a binary question about the data distribution. Our goal is to either accept a **null hypothesis** H_0 (which specifies something about this distribution) or to reject it in favor of an **alternative hypothesis** H_1.

If H_0 (similarly H_1) completely specifies the probability distribution for the data, then the hypothesis is **simple**. Otherwise it is **composite**.

Today we'll focus on testing simple null hypotheses H_0.

Simple vs. composite

Example: Let $X_1, \ldots, X_6$ be the number of times we obtain 1 to 6 in n dice rolls. This null hypothesis is simple:

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

Simple vs. composite

Example: Let $X_1, \ldots, X_6$ be the number of times we obtain 1 to 6 in n dice rolls. This null hypothesis is simple:

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

We might wish to test this null hypothesis against the simple alternative hypothesis

$$H_1 : (X_1, \ldots, X_6) \sim \text{Multinomial}\left(n, \left(\tfrac{1}{9}, \tfrac{1}{9}, \tfrac{1}{9}, \tfrac{2}{9}, \tfrac{2}{9}, \tfrac{2}{9}\right)\right),$$

or perhaps against the compositive alternative hypothesis

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

Simple vs. composite

Example: Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$ be the displacement vectors $P_{30} - P_0, P_{60} - P_{30}, P_{90} - P_{60}, \ldots$ where $P_t \in \mathbb{R}^2$ is the position of a particle at time t in Perrin's experiment. Einstein's theory corresponds to the simple null hypothesis

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}(0, 2.23 \times 10^{-7} I).$$

To test the theory qualitatively, but possibly allow for an error in Einstein's formula for σ^2, we might test the composite null hypothesis

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}(0, \sigma^2 I) \text{ for some } \sigma^2 > 0.$$

One can pose a number of different possible alternative hypotheses H_1 to the above nulls.

Test statistics

A **test statistic** $T := T(X_1, \ldots, X_n)$ is any statistic such that extreme values (large or small) of T provide evidence against H_0.

Example: Let $X_1, \ldots, X_6$ count the results from n dice rolls, and let

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

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

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

Test statistics

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

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}(0, 2.23 \times 10^{-7} I).$$

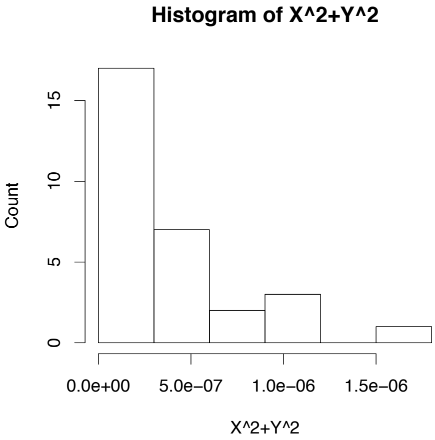
the following are possible test statistics:

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$
$$\bar{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n)$$
$$V = \frac{1}{n}(X_1^2 + Y_1^2 + \ldots + X_n^2 + Y_n^2)$$

(Values of $\bar{X}$ or $\bar{Y}$ much larger or smaller than 0, or values of V much larger or smaller than $2 \times 2.23 \times 10^{-7}$, provide evidence against H_0 in favor of various alternatives H_1.)

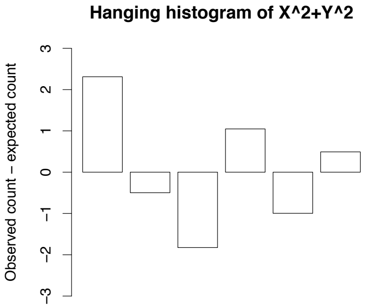
Test statistics from histograms

Let $R_i = X_i^2 + Y_i^2$. Suppose we are interested in testing whether $R_1, \ldots, R_n$ are distributed as $2.23 \times 10^{-7}\chi_2^2$ (their distribution under H_0). We can plot a histogram of these values:



Test statistics from histograms

Deviations from $2.23 \times 10^{-7}\chi^2_2$ are better visualized by a hanging histogram, which plots $O_i - E_i$ where O_i is the observed count for bin i and E_i is the expected count under the $2.23 \times 10^{-7}\chi^2_2$ distribution:



A test statistic can be $T = \sum_{i=1}^{6}(O_i - E_i)^2$.

Test statistics from histograms

Problem: Let p_i be the probability that the hypothesized chi-squared distribution assigns to bin i. If H_0 were true, then $O_i \sim \text{Binomial}(n, p_i)$ and $E_i = np_i = \mathbb{E}[O_i]$. So

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

The variation in O_i is smaller, and scales approximately linearly with p_i, if p_i is close to 0. This might explain why the bars were smaller on the right side of the hanging histogram.

Test statistics from histograms

Problem: Let p_i be the probability that the hypothesized chi-squared distribution assigns to bin i. If H_0 were true, then $O_i \sim \text{Binomial}(n, p_i)$ and $E_i = np_i = \mathbb{E}[O_i]$. So

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

The variation in O_i is smaller, and scales approximately linearly with p_i, if p_i is close to 0. This might explain why the bars were smaller on the right side of the hanging histogram.

Solution: We can "stabilize the variance" by looking at $\frac{O_i - E_i}{\sqrt{E_i}} = \frac{O_i - E_i}{\sqrt{np_i}}$.

Test statistics from histograms

Problem: Let p_i be the probability that the hypothesized chi-squared distribution assigns to bin i. If H_0 were true, then $O_i \sim \text{Binomial}(n, p_i)$ and $E_i = np_i = \mathbb{E}[O_i]$. So

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

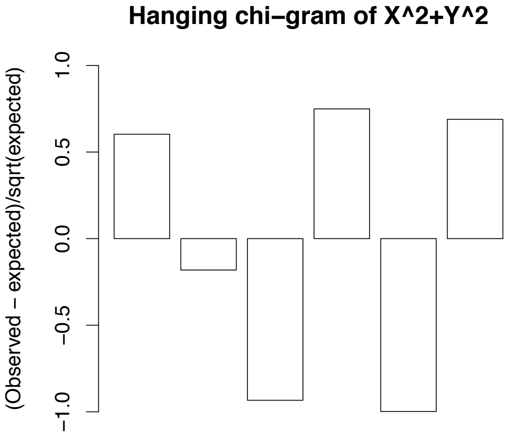
The variation in O_i is smaller, and scales approximately linearly with p_i, if p_i is close to 0. This might explain why the bars were smaller on the right side of the hanging histogram.

Solution: We can "stabilize the variance" by looking at $\frac{O_i - E_i}{\sqrt{E_i}} = \frac{O_i - E_i}{\sqrt{np_i}}$.

Or alternatively, we can look at $\sqrt{O_i} - \sqrt{E_i}$. (Taylor expansion of $\sqrt{x}$ around $x = E_i$ yields $\sqrt{O_i} - \sqrt{E_i} \approx \frac{1}{2\sqrt{E_i}}(O_i - E_i)$, so this has a similar effect as $\frac{O_i - E_i}{2\sqrt{E_i}}$ when $O_i - E_i$ is small.)

Test statistics from histograms

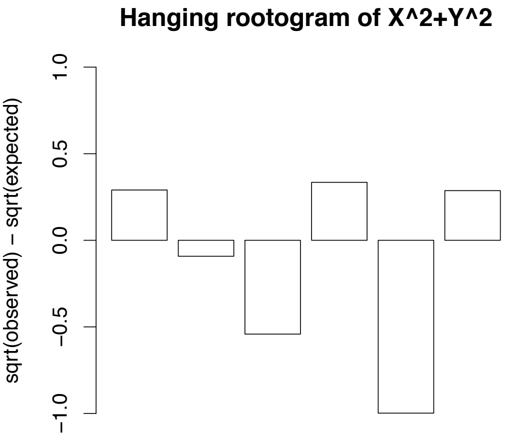
The hanging chi-gram plots $\frac{O_i - E_i}{\sqrt{E_i}}$:



The test statistic $T = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$ is called **Pearson's chi-squared statistic for goodness of fit**.

Test statistics from histograms

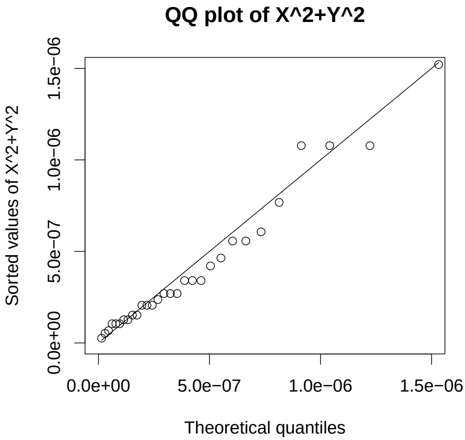
Tukey's hanging rootogram plots $\sqrt{O_i} - \sqrt{E_i}$:



We may take as test statistic $T = \sum_{i=1}^{6} (\sqrt{O_i} - \sqrt{E_i})^2$.

Test statistics from QQ plots

A **QQ plot** (or probability plot) compares the sorted values of $R_1, \ldots, R_n$ with the $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$ quantiles of the hypothesized $2.23 \times 10^{-7}\chi_2^2$ distribution:



Values close to the line $y = x$ indicate a good fit.

Test statistics from QQ plots

How do we get a test statistic from a QQ plot? One way is to take the maximum vertical deviation from the $y = x$ line: Let $R_{(1)} < \ldots < R_{(n)}$ be the sorted values of $R_1, \ldots, R_n$. Take

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

where F is the CDF of the $2.23 \times 10^{-7} \chi^2_2$ distribution so $F^{-1}(t)$ is its t^{th} quantile.

Test statistics from QQ plots

How do we get a test statistic from a QQ plot? One way is to take the maximum vertical deviation from the $y = x$ line: Let $R_{(1)} < \ldots < R_{(n)}$ be the sorted values of $R_1, \ldots, R_n$. Take

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

where F is the CDF of the $2.23 \times 10^{-7}\chi^2_2$ distribution so $F^{-1}(t)$ is its t^{th} quantile.

Problem: For values of R where the distribution has high density, the quantiles are closer together, so we expect a smaller vertical deviation. This explains why we see more vertical deviation in the upper right of the last QQ plot.

Test statistics from QQ plots

Solution: We may stabilize the spacings between quantiles by considering instead

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

Test statistics from QQ plots

Solution: We may stabilize the spacings between quantiles by considering instead

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

This is almost the same as the **one-sample Kolmogorov-Smirnov (K-S) statistic**,

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

(You can show $\frac{i-1}{n} < \frac{i}{n+1} < \frac{i}{n}$, and the difference between T and T_{KS} is negligible for large n.)

Null distributions and type I error

Supposing that we've picked our test statistic T, how large (or small) does T need to be, before we can safely assert that H_0 is false?

Null distributions and type I error

Supposing that we've picked our test statistic T, how large (or small) does T need to be, before we can safely assert that H_0 is false?

In most cases we can never be 100% sure that H_0 is false. But we can compute T from the observed data and compare with the sampling distribution of T if H_0 were true. This is called the **null distribution** of T.

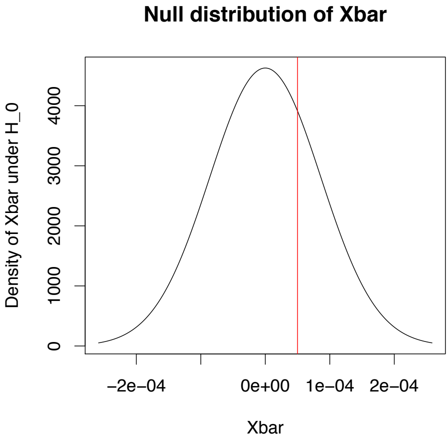
Example: Consider

$$H_0 : (X_1, Y_1), \ldots, (X_n, Y_n) \overset{IID}{\sim} \mathcal{N}(0, 2.23 \times 10^{-7} I).$$

Under H_0, $\bar{X} \sim \mathcal{N}(0, 2.23 \times 10^{-7}/n)$. This normal distribution is the null distribution of $\bar{X}$.

Null distributions and type I error

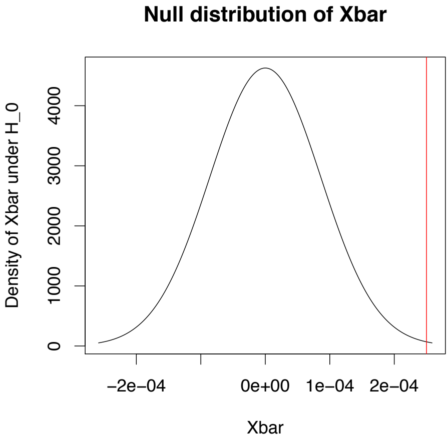
Here's the PDF for the null distribution of $\bar{X}$, when $n = 30$:



If, for the observed data, $\bar{X} = 0.5 \times 10^{-4}$, this would not provide strong evidence against H_0. In this case we might accept H_0.

Null distributions and type I error

Here's the PDF for the null distribution of $\bar{X}$, when $n = 30$:



If, for the observed data, $\bar{X} = 2.5 \times 10^{-4}$, this would provide strong evidence against H_0. In this case we might reject H_0.

Null distributions and type I error

The **rejection region** is the set of values of T for which we choose to reject H_0. The **acceptance region** is the set of values of T for which we choose to accept H_0.

We choose the rejection region so as to control the probability of **type I error**:

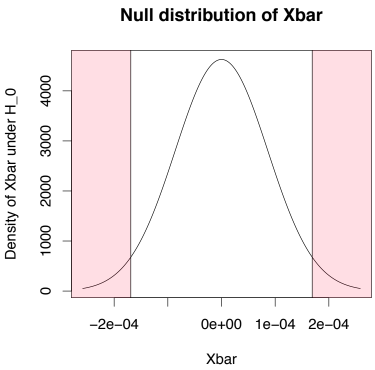
$$\alpha = \mathbb{P}_{H_0}[\text{reject } H_0]$$

This value α is also called the **significance level** of the test.

If, under its null distribution, T belongs to the rejection region with probability α, then the test is level-α.

(Notation: For a simple null hypothesis H_0, we write $\mathbb{P}_{H_0}[\mathcal{E}]$ to denote the probability of event $\mathcal{E}$ under H_0, i.e. the probability of $\mathcal{E}$ if H_0 were true.)

Null distributions and type I error



Example: A (two-sided) level-α test might reject H_0 when $\bar{X}$ falls in the above shaded regions. Mathematically, let $z(\alpha)$ denote the $1-\alpha$ quantile, or "upper α point", of the distribution $\mathcal{N}(0,1)$. As $\bar{X} \sim \mathcal{N}(0, \sigma^2/n)$ under H_0 (where $\sigma^2 = 2.23 \times 10^{-7}$), the rejection region should be $(-\infty, -\frac{\sigma}{\sqrt{n}} \times z(\alpha/2)] \cup [\frac{\sigma}{\sqrt{n}} \times z(\alpha/2), \infty)$.

P-values

The ***p*-value** is the smallest significance level at which your test would have rejected H_0.

For a one-sided test that rejects for large T, letting t_{obs} denote the value of T computed from the observed data, the p-value is $\mathbb{P}_{H_0}[T \geq t_{\text{obs}}]$.

For a two-sided test that rejects at the $\alpha/2$ and $1 - \alpha/2$ quantiles of the null distribution of T, the p-value is 2 times the smaller of $\mathbb{P}_{H_0}[T \geq t_{\text{obs}}]$ and $\mathbb{P}_{H_0}[T \leq t_{\text{obs}}]$.

The p-value provides a quantitative measure of the extent to which the data supports (or does not support) H_0. It is preferable to report the exact p-value, rather than to just say "we rejected at level-0.05".

A word of caution

Accepting (or failing to reject) H_0 **does not** imply there is strong evidence that H_0 is true. Both of the following are possible:

- The particular test statistic you chose is not good at distinguishing the null hypothesis H_0 from the true distribution. Or equivalently, the true distribution is not well-captured by the alternative H_1 that your test statistic is targeting. (For example, in Perrin's data, if there is significant drift in the y direction, you would not detect this using the test statistic $\bar{X}$.)
- You do not have enough data to reject H_0 at the significance level that you desire. In this case your study might be **underpowered**—we'll discuss this issue a couple weeks from now.

Determining the null distribution

To figure out the rejection region, we must understand the null distribution of the test statistic. There are three methods:

- Sometimes we can derive the null distribution exactly, for example in the previous slides where the test statistic is $\bar{X}$ and $X_1, \ldots, X_n$ are normally distributed under H_0.
- Sometimes we can derive an asymptotic approximation, using tools such as the CLT and continuous mapping theorem.
- When H_0 is simple, we can always obtain the null distribution by simulation.

Using an asymptotic null distribution

Example: Let $(X_1, \ldots, X_6)$ denote the counts of 1 to 6 from n rolls of a die, and consider testing the simple null of a fair die

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

using the test statistic

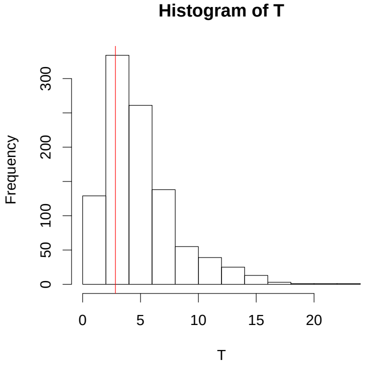
$$T = \left(\frac{X_1}{n} - \frac{1}{6}\right)^2 + \ldots + \left(\frac{X_6}{n} - \frac{1}{6}\right)^2.$$

Recall from last lecture that for large n, T is approximately distributed as $\frac{1}{6n}\chi^2_5$.

To perform an **asymptotic level-$\boldsymbol{\alpha}$ test**, we may reject H_0 when t_{obs} exceeds $\frac{1}{6n}\chi^2_5(\alpha)$, where $\chi^2_n(\alpha)$ denotes the $1-\alpha$ quantile, or "upper α point", of the χ^2_n distribution.

Using a simulated null distribution

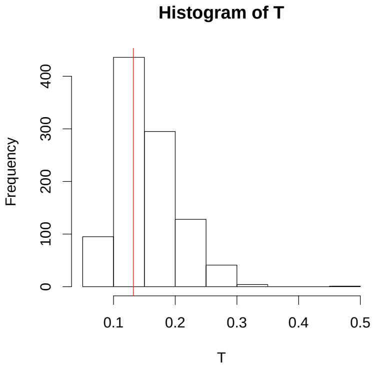
Example: Let T be Pearson's chi-squared statistic for goodness of fit for the values $X_1^2 + Y_1^2, \ldots, X_{30}^2 + Y_{30}^2$ from Perrin's experiments, discussed previously. We may simulate the null distribution of T:



This shows the 1000 values of T across 1000 simulations. The observed value $t_{obs} = 2.83$ for Perrin's real data is in red.

Using a simulated null distribution

Example: Let T be the K-S statistic for $X_1^2 + Y_1^2, \ldots, X_{30}^2 + Y_{30}^2$, discussed previously. We may simulate the null distribution of T:



The observed value $t_{\text{obs}} = 0.132$ for Perrin's real data is in red.

Using a simulated null distribution

We obtain an approximate p-value as the fraction of simulated values of T larger than t_{obs}. (For a two-sided test, we would take either the fraction of simulated values of T larger than t_{obs} or smaller than t_{obs}, and multiply this by 2.)

For Perrin's data, the Pearson chi-squared p-value is 0.754, and the K-S p-value is 0.612. We accept H_0 in both cases, and neither test provides significant evidence against Einstein's theory of Brownian motion.