Chapter 9: Testing Hypotheses and Assessing Goodness of Fit

Donghyun Ko

May 27, 2026

This note follows the chapter-level structure of Rice, Chapter 9, while expanding the lecture material into a more textbook-style exposition. The emphasis is on explanation rather than mere formula collection: we carefully define the language of testing, explain the logic of Type I and Type II errors, develop rejection regions and p-values, prove the Neyman–Pearson lemma and the duality between tests and confidence intervals, introduce generalized likelihood ratio tests, and conclude with model assessment tools such as probability plots and tests for normality. Representative examples are solved in full detail so that an undergraduate reader can follow not only *what* to do, but also *why* each step is taken.

Contents

1 Introduction

1.1 Why hypothesis testing is needed

In statistical work, we are often asked to make claims about a population distribution / population parameter, or a stochastic mechanism that generated the data. However, we almost never observe the entire population. Instead, we see only a sample, and the sample is subject to randomness. Because of this randomness, even when a null claim is true, the sample may look somewhat unusual. Conversely, even when an alternative claim is true, the sample may fail to show a dramatic difference.

A statistical hypothesis is an assumption (claim, conjecture) about the value of a population parameter, and a hypothesis testing is the process of testing the validity of the statistical hypothesis based on a random sample drawn from the population. **In other words, hypothesis testing provides a principled framework for deciding whether the observed data are sufficiently inconsistent with a baseline claim that we should reject it.** Typical questions include:

- "Unmarried workers are more likely to be absent from work than married workers."
- "Vaccinated people are less likely to get infected with COVID than non-vaccinated individuals."

Definition

A **statistical hypothesis** is a claim about a population distribution or about parameters of that distribution.

Hypothesis testing is the formal process of using sample data to evaluate a null claim against an alternative claim.

1.2 The basic language of testing

We write

$$H_0 \quad \text{for the null hypothesis,} \qquad H_A \text{ or } H_1 \quad \text{for the alternative hypothesis.}$$

Usually:

- H_0 is the default, baseline, or status quo claim;
- H_A is the competing scientific or practical claim.

Example

Suppose a neurologist knows that the mean response time without treatment is 1.2 seconds and claims that a new drug *decreases* response time. Then an appropriate formulation is

$$H_0 : \mu = 1.2 \qquad \text{vs} \qquad H_A : \mu < 1.2.$$

The null says there is no decrease from the known baseline, while the alternative says the mean is smaller.

We cannot prove that H_0 is true BUT we can prove that H_0 is much more plausible than H_A given the data! A test never truly proves the null hypothesis. The two possible conclusions are:

- reject H_0,
- fail to reject H_0.

The phrase "fail to reject" is intentionally cautious. It means the data did not provide enough evidence against the null; it does *not* mean the null has been established as true.

Example

Rice begins with a very simple example that captures the essential idea. Suppose there are two possible coins:

- Coin 0: probability of heads is 0.5,
- Coin 1: probability of heads is 0.7.

One of the two coins is chosen, tossed 10 times, and only the number of heads is observed. Let X be the number of heads. Then, the two competing hypotheses are

$$H_0 : \ X \sim \text{Binomial}(10, 0.5), \qquad H_1 : \ X \sim \text{Binomial}(10, 0.7).$$

This setting is artificial, but it reveals the central structure of testing:

- There are two competing explanations;
- The data are random under each explanation;
- We need a rule for deciding which explanation is better supported.

Suppose we observe $X = 2$. Then,

$$\frac{P_0(X = 2)}{P_1(X = 2)} \approx \frac{0.0439}{0.0014} \approx 30.$$

So observing 2 heads is about 30 times more likely under the fair coin than under the coin with head probability 0.7. If instead we observe $X = 8$, then

$$\frac{P_0(X = 8)}{P_1(X = 8)} \approx \frac{0.0439}{0.2335} \approx 0.188.$$

This strongly favors the second coin. The **likelihood ratio** tells us how much more plausible one hypothesis is than another in light of the observed data. This is the unifying idea behind much of Chapter 9.

2 Basic Elements of Hypothesis Testing

2.1 Decision tools and the logic of testing

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta),$$

where θ is an unknown parameter. In the simplest testing setup, we compare two specific possibilities:

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta = \theta_A$$

These are called **simple hypotheses** because each hypothesis specifies a single value of the parameter. Once data $x = (x_1, \ldots, x_n)$ are observed, the goal is to decide which of the two hypotheses is better supported by the sample.

Definition

In hypothesis testing, we use the observed sample to decide between two competing claims:

$$H_0 \quad \text{and} \quad H_A$$

The decision is made using a formal tool such as a **Bayes factor, a test statistic, a rejection region, or a p-value**, which are all going to be discussed in the next steps.

There are only two possible conclusions:

- reject H_0 in favor of H_A;
- fail to reject H_0.

The second conclusion must be interpreted carefully. If we fail to reject H_0, this does *not* mean that H_0 has been proved true. It means only that the data were not sufficiently incompatible with H_0.

Remark

A hypothesis test is not a proof machine. It is a decision procedure under uncertainty. Because the sample is random, even false hypotheses may occasionally look plausible, and even true alternatives may sometimes fail to produce dramatic evidence.

2.2 Bayesian perspective: Bayes factors

In the Bayesian view, neither hypothesis is automatically treated as the "status quo." Instead, both hypotheses are assigned prior support, and the data update that support. Let the prior probabilities at θ_0 and θ_A be $f_\Theta(\theta_0)$ and $f_\Theta(\theta_A)$. After observing data x, the posterior support for the two hypotheses is proportional to

$$f(x \mid \theta_0) f_\Theta(\theta_0) \qquad \text{and} \qquad f(x \mid \theta_A) f_\Theta(\theta_A).$$

Definition

The **Bayes factor** comparing H_0 to H_A is

$$BF = \frac{f_{\Theta|X}(\theta_0 \mid x)}{f_{\Theta|X}(\theta_A \mid x)} = \frac{f_\Theta(\theta_0)}{f_\Theta(\theta_A)} \cdot \frac{f(x \mid \theta_0)}{f(x \mid \theta_A)}.$$

Thus, the Bayes factor is the product of a **prior ratio** and a **likelihood ratio**.

Interpretation:

- $BF > 1$ suggests more support for H_0;
- $BF \leq 1$ suggests more support for H_A;

A commonly used rule of thumb is:

$$\frac{1}{3} \leq BF < 1 \quad \Longrightarrow \quad \text{substantial support for } H_A,$$

$$BF < \frac{1}{3} \quad \Longrightarrow \quad \text{strong support for } H_A.$$

Intuition

The Bayes factor answers the question: *After combining prior information and the observed data, which hypothesis is better supported?* It is a direct comparison of the two hypotheses.

Example

Daily notification app example Let X be the number of days (out of 6) on which notifications are received.

$$H_0 : X \sim \text{Binomial}(6, 0.5), \qquad H_A : X \sim \text{Binomial}(6, 0.8).$$

Assume the two apps are equally likely a priori, so

$$f_\Theta(\theta_0) = f_\Theta(\theta_A).$$

Hence the prior ratio is 1, and the Bayes factor becomes

$$BF(x) = \frac{P_{\theta_0}(X = x)}{P_{\theta_A}(X = x)}.$$

Since

$$P_{\theta_0}(X = x) = \binom{6}{x}(0.5)^x(0.5)^{6-x}$$

and

$$P_{\theta_A}(X = x) = \binom{6}{x}(0.8)^x(0.2)^{6-x},$$

we have

$$BF(x) = \frac{\binom{6}{x}(0.5)^x(0.5)^{6-x}}{\binom{6}{x}(0.8)^x(0.2)^{6-x}} = \frac{(0.5)^6}{(0.8)^x(0.2)^{6-x}}.$$

(a) Calculate the Bayes factor for each outcome.
Using the formula above:

$$BF(0) = \frac{(0.5)^6}{(0.2)^6} \approx 244.14, \qquad BF(1) = \frac{(0.5)^6}{(0.8)(0.2)^5} \approx 61.04,$$

$$BF(2) = \frac{(0.5)^6}{(0.8)^2(0.2)^4} \approx 15.26, \qquad BF(3) = \frac{(0.5)^6}{(0.8)^3(0.2)^3} \approx 3.81,$$

$$BF(4) = \frac{(0.5)^6}{(0.8)^4(0.2)^2} \approx 0.95, \qquad BF(5) = \frac{(0.5)^6}{(0.8)^5(0.2)} \approx 0.24,$$

$$BF(6) = \frac{(0.5)^6}{(0.8)^6} \approx 0.06.$$

So,

x	0	1	2	3	4	5	6
$BF(x)$	244.14	61.04	15.26	3.81	0.95	0.24	0.06

Recall:

- $BF(x) > 1$ means the outcome supports H_0;
- $BF(x) < 1/3$ means the outcome strongly supports H_A.

(b) Find the rejection region.
By definition, Rejection Region (RR) is a set of outcomes that support H_A strongly with $BF < 1/3$ such that

$$RR = \{x : BF(x) < 1/3\}.$$

From the table, this happens only for $x = 5$ and $x = 6$. Therefore,

$$\boxed{RR = \{5, 6\}.}$$

So we reject H_0 when the observed number of notification days is either 5 or 6.

(c) Determine the outcomes that support H_0.
These are the outcomes for which $BF(x) > 1$. From the table,

$$\boxed{\{0, 1, 2, 3\}}$$

support H_0. Note that when $x = 4$, we have $BF(4) \approx 0.95 < 1$, so it slightly favors H_A, but not strongly enough to fall in the rejection region.

(d)(i) Compute P_{θ_0}(Reject H_0).
Under H_0 is true, we reject only when $X = 5$ or $X = 6$. Thus

$$P_{\theta_0}(\text{Reject } H_0) = P_{\theta_0}(X = 5) + P_{\theta_0}(X = 6).$$

Since under H_0, $X \sim \text{Binomial}(6, 0.5)$,

$$P_{\theta_0}(X = 5) = \binom{6}{5}(0.5)^5(0.5)^1 = \binom{6}{5}(0.5)^6 = \frac{6}{64},$$

and

$$P_{\theta_0}(X = 6) = \binom{6}{6}(0.5)^6 = \frac{1}{64}.$$

Therefore,

$$P_{\theta_0}(\text{Reject } H_0) = \frac{6}{64} + \frac{1}{64} = \frac{7}{64} \approx 0.109.$$

(d)(ii) Compute P_{θ_A}(Fail to reject H_0).
Under H_A is true, a failure to reject occurs when $X \notin \{5, 6\}$, so

$$P_{\theta_A}(\text{Fail to reject } H_0) = 1 - P_{\theta_A}(X = 5) - P_{\theta_A}(X = 6).$$

Now under H_A, $X \sim \text{Binomial}(6, 0.8)$, so

$$P_{\theta_A}(X = 5) = \binom{6}{5}(0.8)^5(0.2) = 6(0.8)^5(0.2) \approx 0.3932,$$

and

$$P_{\theta_A}(X = 6) = (0.8)^6 \approx 0.2621.$$

Hence

$$P_{\theta_A}(\text{Fail to reject } H_0) = 1 - (0.3932 + 0.2621) \approx 0.3447.$$

Summary.

- Rejection region: $\boxed{\{5, 6\}}$
- Outcomes supporting H_0: $\boxed{\{0, 1, 2, 3\}}$
- $P_{\theta_0}(\text{Reject } H_0) \approx 0.109$
- $P_{\theta_A}(\text{Fail to reject } H_0) \approx 0.345$

2.3 Frequentist perspective: two types of error

In the frequentist approach, H_0 is treated as the baseline or status-quo claim. The test is designed to challenge H_0 and reject it only when the evidence is sufficiently strong. Because decisions are made from random data, two mistakes are possible.

Definition

- A **Type I error** occurs when we reject H_0 even though H_0 is true.

  $$\alpha = P_{H_0}(\text{Reject } H_0).$$

  The quantity α is called the **significance level** of the test. It is also called the **false positive rate**.

- A **Type II error** occurs when we fail to reject H_0 even though H_A is true.

  $$\beta = P_{H_A}(\text{Fail to reject } H_0).$$

  This is also called the **false negative rate**.

- The **power** of the test is the probability of correctly rejecting H_0 when H_A is true:

  $$\text{Power} = P_{H_A}(\text{Reject } H_0) = 1 - \beta.$$

Intuition

A good test should have

- small α (few false alarms), and
- large power (strong ability to detect false null hypotheses).

In practice, there is usually a trade-off: if we make rejection very difficult in order to reduce α, then we often increase β.

The answer for which error is worse depends on the scientific context. A standard illustration is the justice system:

$$H_0 : \text{person is innocent}, \qquad H_A : \text{person is guilty}.$$

Then:

- a Type I error means convicting an innocent person;
- a Type II error means failing to convict a guilty person.

For most serious crimes, society usually regards the Type I error as worse. That is why the legal system sets a high standard of proof before rejecting the null claim of innocence.

Remark

This analogy helps explain why the null hypothesis is often treated as the claim we are especially cautious about rejecting.

2.4 Frequentist perspective: rejection regions

A **test statistic** is a function of the data used to make the testing decision. It is usually denoted by

$$T(X), \qquad \text{where } X = (X_1, \ldots, X_n).$$

Definition

A **rejection region** (RR) is the set of values of the test statistic for which we reject H_0. Its complement is called the **acceptance region**:

$$AR = RR^c.$$

For a level-α test, the rejection region is chosen so that under H_0,

$$P_{H_0}\{T(X) \in RR\} = \alpha$$

or at least

$$P_{H_0}\{T(X) \in RR\} \leq \alpha.$$

Thus, the frequentist procedure works in two steps:

1. Determine a rejection region using the distribution of the test statistic under H_0;
2. Check whether the observed test statistic falls inside that region.

If it does, reject H_0. If it does not, fail to reject H_0.

Example

RDU airport example RDU airport wants to know whether the mean number of passengers waiting per day has increased this year. Historical data suggest that the old mean is 205, and the population standard deviation is known to be $\sigma = 107$. Let

$$Y_1, \ldots, Y_{31} \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \qquad \sigma = 107.$$

We want to test

$$H_0 : \mu = 205 \qquad \text{vs} \qquad H_A : \mu > 205.$$

Step 1. What statistic should we use? To test whether the mean has increased, a natural statistic is the sample mean

$$\bar{Y} = \frac{1}{31} \sum_{i=1}^{31} Y_i.$$

Since the alternative is $H_A : \mu > 205$, *large* values of $\bar{Y}$ provide evidence for H_A. It is also convenient to standardize $\bar{Y}$ and use

$$Z = \frac{\bar{Y} - 205}{107/\sqrt{31}}.$$

Both $\bar{Y}$ and Z contain the same information, but Z is easier to use, because under H_0, it has the standard normal distribution.

Step 2. Which values support the alternative?
Because the alternative says the true mean is larger than 205, larger values of

$$\bar{Y} \quad \text{or equivalently} \quad Z$$

support H_A. So the rejection region should be of the form

$$RR = \{\bar{Y} > c_1\} \qquad \text{or equivalently} \qquad RR = \{Z > c_2\},$$

for some cutoff values c_1, c_2.

Step 3. Find the rejection region with Type I error rate 1%.
We want the test to have significance level

$$\alpha = 0.01.$$

That means we choose the cutoff so that

$$P_{H_0}(\text{Reject } H_0) = 0.01.$$

Under H_0,

$$\bar{Y} \sim N\left(205, \frac{107^2}{31}\right),$$

so

$$Z = \frac{\bar{Y} - 205}{107/\sqrt{31}} \sim N(0,1).$$

Therefore we choose c_2 so that

$$P(Z > c_2) = 0.01, \qquad Z \sim N(0,1).$$

From the standard normal table,

$$c_2 \approx 2.33.$$

Hence, using the standardized statistic,

$$RR = \{z : z > 2.33\}.$$

Now convert this back to $\bar{Y}$:

$$\frac{\bar{Y} - 205}{107/\sqrt{31}} > 2.33$$

which implies

$$\bar{Y} > 205 + 2.33\frac{107}{\sqrt{31}}.$$

Numerically,

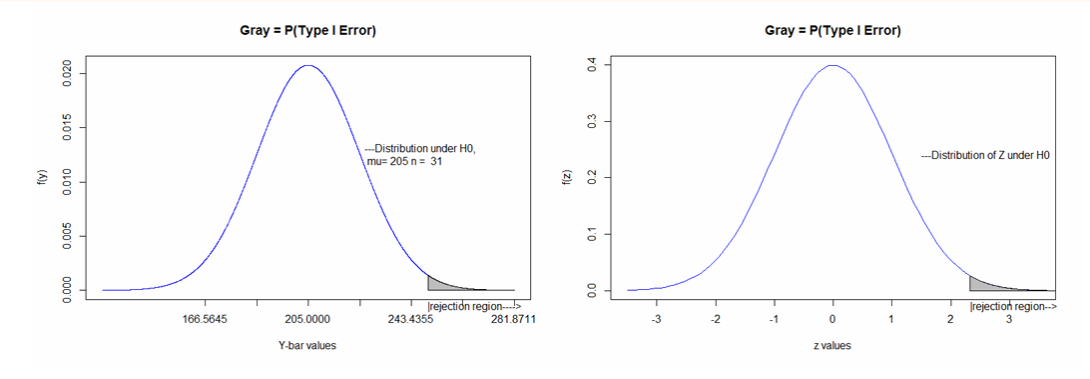
$$205 + 2.33\frac{107}{\sqrt{31}} \approx 249.7.$$

So the rejection region can also be written as

$$\boxed{RR = \{\bar{Y} > 249.7\}}$$

and the acceptance region is

$$\boxed{AR = \{\bar{Y} \leq 249.7\}.}$$



Interpretation. If the observed sample mean is greater than 249.7, we reject H_0 and conclude that there is statistically significant evidence that the mean waiting count has increased. If the

observed sample mean is at most 249.7, we fail to reject H_0. Also, because the rejection region was chosen so that

$$P_{H_0}(\bar{Y} > 249.7) = 0.01,$$

we will falsely reject H_0 only 1% of the time when H_0 is true.

Step 4. Calculate the Type II error when the true mean is $\mu_A = 275$. Now suppose the true mean is actually

$$\mu_A = 275.$$

Then

$$\bar{Y} \sim N\left(275, \frac{107^2}{31}\right).$$

A Type II error means *failing to reject H_0 even though the alternative is true.* Since the acceptance region is $\bar{Y} \leq 249.7$, we get

$$\beta(275) = P(\bar{Y} \leq 249.7 \mid \mu = 275).$$

Standardizing under the true mean $\mu = 275$,

$$\beta(275) = P\left(Z \leq \frac{249.7 - 275}{107/\sqrt{31}}\right).$$

Compute the standardized value:

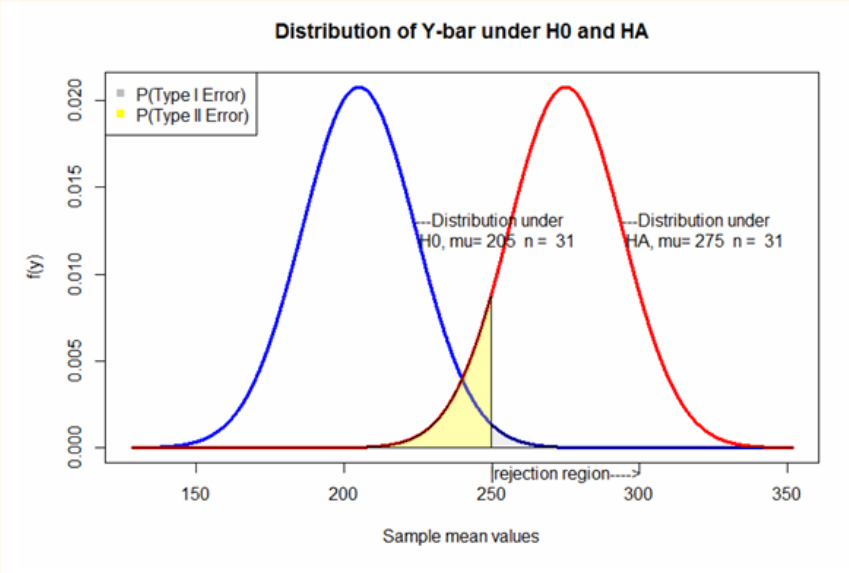
$$\frac{249.7 - 275}{107/\sqrt{31}} \approx -1.32.$$

Therefore,

$$\beta(275) \approx P(Z \leq -1.32) \approx 0.09.$$

So the Type II error probability at $\mu_A = 275$ is

$$\boxed{\beta(275) \approx 0.09.}$$



Step 5. Calculate the power.
The power of a test at μ_A is

$$\text{Power}(\mu_A) = 1 - \beta(\mu_A) = P_{H_A}(\text{Reject } H_0).$$

Thus, at $\mu_A = 275$,

$$\text{Power}(275) = 1 - \beta(275) \approx 1 - 0.09 = 0.91.$$

Hence,

$$\boxed{\text{Power}(275) \approx 0.91.}$$

Final remarks.

- Type I error is controlled at $\alpha = 0.01$.
- At $\mu_A = 275$, the Type II error is about $\beta(275) \approx 0.09$.
- Therefore, the power is about 0.91

In general, the farther the true mean μ_A is above the null value 205, the smaller the Type II error and the larger the power. For a fixed Type I error rate α, one of the best ways to improve power is to increase the sample size n.

Example

Let

$$Y \sim \text{Uniform}(\theta, \theta + 1),$$

and consider

$$H_0 : \theta = 0 \qquad \text{vs} \qquad H_A : \theta > 0.$$

Suppose the rejection region is given by

$$RR = \{y : y > 0.95\}.$$

That is, we reject H_0 whenever the observed value satisfies $y > 0.95$.

(i) Type I error rate. The Type I error rate is

$$\alpha = P_{\theta=0}(\text{Reject } H_0) = P_{\theta=0}(Y > 0.95).$$

Under H_0, we have

$$Y \sim \text{Uniform}(0, 1).$$

Therefore,

$$P_{\theta=0}(Y > 0.95) = 1 - 0.95 = 0.05.$$

So the Type I error rate is

$$\boxed{\alpha = 0.05.}$$

(ii) Calculate $\beta(\theta_A = 0.5)$ and $PWR(\theta_A = 0.5)$. When $\theta_A = 0.5$, we have

$$Y \sim \text{Uniform}(0.5, 1.5).$$

A Type II error means *failing to reject* H_0 when the alternative is true. Since we reject when $Y > 0.95$, we fail to reject when $Y \leq 0.95$. Hence

$$\beta(0.5) = P_{\theta=0.5}(Y \leq 0.95).$$

Because $Y \sim \text{Uniform}(0.5, 1.5)$, this probability is the length of the interval [0.5, 0.95] divided by the total length of the support [0.5, 1.5]. So, the type 2 error ($\beta(\theta_A = 0.5)$) is

$$\boxed{\beta(0.5) = \frac{0.95 - 0.5}{1.5 - 0.5} = \frac{0.45}{1} = 0.45}$$

The power ($PWR(\theta_A = 0.5)$) is

$$\boxed{PWR(0.5) = 1 - \beta(0.5) = 1 - 0.45 = 0.55}$$

2.5 Frequentist perspective: p-value

Another way to carry out hypothesis testing is through the **p-value**. Instead of explicitly specifying a rejection region first, we compute a probability that measures how extreme the observed test statistic is under the null hypothesis.

Definition

The **p-value** is the probability, computed under H_0, of observing a test statistic as extreme as or more extreme than the one actually observed.

General setting. Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta),$$

and we want to test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta > \theta_0.$$

Let $T = T(X)$ be a test statistic.

Testing procedure.

1. Compute the observed value of the test statistic:

$$T_{\text{obs}} = T(x).$$

2. Compute the p-value:

$$\text{p-value} = P_{H_0}\big(T(X) > T_{\text{obs}}\big).$$

   This requires knowledge of the distribution of $T(X)$ under H_0.

Intuition

For a right-tailed test, "more extreme" means *larger* values of the test statistic. For a left-tailed test, it means *smaller* values. For a two-sided test, it means values farther from the null value in either direction.

For example, if

$$X_i \sim N(\theta, 5^2),$$

and we use

$$T = \frac{\bar{X} - \theta_0}{5/\sqrt{n}},$$

then the p-value for testing $H_0 : \theta = \theta_0$ versus $H_A : \theta > \theta_0$ is

$$P_{H_0}(T > T_{\text{obs}}).$$

2.5.1 Two interpretations of the p-value

There are two commonly used interpretations of the p-value.

Definition

The p-value is the probability of obtaining results as extreme as or more extreme than the observed result, assuming that the null hypothesis is true. In other words, the p-value is the **smallest significance level** α at which we would reject H_0.

Interpretation.

- A **large p-value** means the observed data are reasonably compatible with H_0, so we fail to reject H_0.
- A **small p-value** means the observed data are not very compatible with H_0, so we reject H_0.

Remark

The p-value does not directly measure evidence *for* the alternative. Rather, it measures how surprising the observed data would be *if the null hypothesis were true.*

Decision rule. If a significance level α is specified in advance, then

$$\text{Reject } H_0 \text{ if p-value} \leq \alpha, \qquad \text{otherwise fail to reject } H_0.$$

Example

Example 1: Drug effectiveness (mean test)
A neurologist claims that a specific drug decreases response time. The known mean response time without treatment is 1.2 sec. A sample of 100 rats injected with the drug yields

$$\bar{X} = 1.05, \qquad s = 0.5.$$

Step 1: Hypotheses

$$H_0 : \mu = 1.2, \qquad H_A : \mu < 1.2.$$

Step 2: Test statistic Because the claim is that the drug *reduces* response time, this is a left-tailed test. A natural test statistic is

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Substituting the observed values,

$$Z_{\text{obs}} = \frac{1.05 - 1.2}{0.5/\sqrt{100}} = \frac{-0.15}{0.05} = -3.$$

For a large sample, under H_0 this statistic is approximately standard normal:

$$Z \approx N(0, 1).$$

Step 3: p-value
Since this is a left-tailed test,

$$\text{p-value} = P(Z < -3) \approx 0.0013.$$

Conclusion
Because the p-value is very small, we reject H_0. There is strong statistical evidence that the drug reduces response time.

Example

Example 2: Proportion test

We want to determine whether the proportion of individuals with a certain trait is less than 0.5. A sample of 100 individuals is observed, and 44 are found to have the trait.

(i) Hypotheses

$$H_0 : p = 0.5, \qquad H_A : p < 0.5.$$

(ii) Test statistic

Let

$$\hat{p} = \frac{44}{100} = 0.44.$$

Use the test statistic

$$T = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}},$$

where $p_0 = 0.5$ and $n = 100$. Under H_0, for large n,

$$T \approx N(0, 1).$$

Now compute the observed value:

$$T_{\text{obs}} = \frac{0.44 - 0.5}{\sqrt{0.5(1-0.5)/100}} = \frac{-0.06}{0.05} = -1.2.$$

p-value

Because the alternative is left-tailed,

$$\text{p-value} = P(Z < -1.2) \approx 0.115.$$

Conclusion

Since the p-value is not small, we fail to reject H_0. There is not enough evidence to conclude that the true proportion is less than 0.5.

2.5.2 Remarks: statistical vs practical significance

Definition

If a hypothesis test rejects H_0, we say the result is **statistically significant**. **Practical significance** refers to whether the magnitude of the effect is large enough to matter in real-world.

This means that the observed data provide evidence that the true parameter differs from the hypothesized value. Equivalently, the observed outcome would be unlikely to occur by random sampling variation alone if H_0 were true. However, statistical significance does *not* necessarily imply that the effect is large or important.

Intuition

A very small effect may become statistically significant if the sample size is large enough. Therefore, when interpreting a test result, we should consider both statistical significance and effect size.

2.5.3 Remarks: issues related to the p-value

Remark

The p-value is one of the most widely used quantities in applied statistics, but it is also one of the most misunderstood.

Common misconceptions include:

- The p-value does **not** equal the probability that H_0 is true.
- The p-value does **not** equal the probability that the data were produced by "random chance alone."
- A small p-value does **not** measure the size of an effect.
- Statistical significance does **not** automatically imply scientific or practical importance.

The correct interpretation is that the p-value measures how compatible the observed data are with the model specified under H_0.

2.5.4 Test statistics

Both the rejection region (RR) approach and the p-value approach depend on a **test statistic**. So far, we have used statistics that were natural for the parameter of interest, such as

- the sample mean for a population mean,
- the sample proportion for a population proportion.

We now introduce a more general framework. Two commonly used test statistics are:

- the **Wald test**,
- the **Likelihood Ratio Test (LRT)**.

2.5.5 Wald test

Suppose we want to test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta > \theta_0.$$

Let $\hat{\theta}$ be an estimator of θ, and suppose that for large n,

$$\hat{\theta} \approx N\big(\theta, \mathrm{Var}_\theta(\hat{\theta})\big).$$

Definition

The **Wald test statistic** is obtained by standardizing $\hat{\theta}$ under the null hypothesis:

$$Z = \frac{\hat{\theta} - \theta_0}{SE_{\theta_0}(\hat{\theta})}, \qquad SE_{\theta_0}(\hat{\theta}) = \sqrt{\mathrm{Var}_{\theta_0}(\hat{\theta})}.$$

If H_0 is true, then for large samples,

$$Z \approx N(0,1).$$

For observed data $x_1, \ldots, x_n$, the observed Wald statistic is

$$z_{\text{obs}} = \frac{\hat{\theta} - \theta_0}{SE_{\theta_0}(\hat{\theta})},$$

where both $\hat{\theta}$ and the standard error are computed from the sample.

Decision rules for the Wald test

Alternative H_A	Rejection Region	p-value
$\theta > \theta_0$	$\{z : z > z_\alpha\}$	$P(Z > z_{\text{obs}})$
$\theta < \theta_0$	$\{z : z < -z_\alpha\}$	$P(Z < z_{\text{obs}})$
$\theta \neq \theta_0$	$\{z : \lvert z\rvert > z_{\alpha/2}\}$	$2P(Z > \lvert z_{\text{obs}}\rvert)$

Thus:

- For a right-tailed test, reject when the observed standardized value is sufficiently large.
- For a left-tailed test, reject when it is sufficiently small.
- For a two-sided test, reject when its magnitude is sufficiently large.

Remark

For a specified significance level α, the RR approach and the p-value approach are equivalent: we reject H_0 whenever the p-value is at most α.

Unknown standard error and large-sample implementation

In practice, the exact standard error under H_0 may be unknown. In that case, we replace it with a consistent estimator:

$$\widehat{SE}(\hat{\theta}).$$

Then the Wald statistic becomes

$$Z = \frac{\hat{\theta} - \theta_0}{\widehat{SE}(\hat{\theta})}.$$

By Slutsky's theorem, under H_0,

$$Z \approx N(0,1) \qquad \text{for large } n.$$

So the Wald testing procedure remains the same, except that we use the estimated standard error.

A natural question is: how do we estimate $\mathrm{Var}(\hat{\theta})$ or $SE(\hat{\theta})$? When $\hat{\theta}$ is the MLE, we often use its asymptotic distribution:

$$\hat{\theta} \approx N\left(\theta, \{nI(\theta)\}^{-1}\right),$$

where $I(\theta)$ is the Fisher information. In practice, this variance is typically estimated by plugging in $\hat{\theta}$, for example using

$$I(\hat{\theta}) \qquad \text{or equivalently} \qquad -\frac{\ell''(\hat{\theta})}{n}.$$

Example

Wald test example 1: Normal mean

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2),$$

where σ^2 is unknown. We want to test

$$H_0 : \theta = 0 \qquad \text{vs} \qquad H_A : \theta > 0.$$

A natural estimator is

$$\hat{\theta} = \bar{X}.$$

Since

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n},$$

we estimate the standard error by

$$\widehat{SE}(\bar{X}) = \frac{S}{\sqrt{n}}.$$

Therefore, the Wald statistic is

$$Z = \frac{\bar{X} - 0}{S/\sqrt{n}} = \frac{\bar{X}}{S/\sqrt{n}}.$$

For a right-tailed test, the rejection rule is:

$$\text{Reject } H_0 \text{ if } Z > z_\alpha,$$

or equivalently if the p-value

$$P(Z > Z_{\text{obs}})$$

is sufficiently small.

Example

Wald test example 2: Bernoulli parameter

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta),$$

and we want to test

$$H_0 : \theta = \frac{3}{4} \qquad \text{vs} \qquad H_A : \theta \neq \frac{3}{4}.$$

A natural estimator is

$$\hat{\theta} = \hat{p}.$$

Since for a Bernoulli random variable,

$$\mathrm{Var}(\hat{p}) = \frac{\theta(1-\theta)}{n},$$

under H_0 the standard error is

$$SE_{H_0}(\hat{p}) = \sqrt{\frac{(3/4)(1/4)}{n}}.$$

Hence, the Wald statistic is

$$Z = \frac{\hat{p} - 3/4}{\sqrt{(3/4)(1/4)/n}}.$$

Under H_0, for large samples,

$$Z \approx N(0,1).$$

Because the alternative is two-sided, we reject H_0 when

$$|Z| > z_{\alpha/2},$$

or equivalently when the p-value

$$2P(Z > |z_{\text{obs}}|)$$

is small.

2.5.6 Likelihood Ratio Test (LRT)

We now introduce another general testing framework based on the likelihood function.

Setting. Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta),$$

and we want to test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta = \theta_A.$$

Both hypotheses are **simple**, since they completely specify the distribution. (If a hypothesis does not fully specify the distribution, it is called **composite**.)

Likelihood function.

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

Definition

The **likelihood ratio statistic** is

$$\Lambda = \frac{L(\theta_0)}{L(\theta_A)}.$$

Decision rule (LRT).

- Reject H_0 if $\Lambda < c$,
- where c is chosen so that

$$P_{H_0}(\Lambda < c) = \alpha.$$

Intuition

Interpretation:

- If Λ is small, then $L(\theta_A)$ is much larger than $L(\theta_0)$, so the data favor H_A.
- If Λ is large, the data favor H_0.

Example

LRT for exponential distribution

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda),$$

with PDF

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad x > 0.$$

We want to test

$$H_0 : \lambda = \lambda_0 \qquad \text{vs} \qquad H_A : \lambda = \lambda_A, \quad \text{where } \lambda_A > \lambda_0.$$

Step 1: Likelihood function

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum X_i}.$$

Step 2: Likelihood ratio

$$\Lambda = \frac{L(\lambda_0)}{L(\lambda_A)} = \frac{\lambda_0^n e^{-\lambda_0 \sum X_i}}{\lambda_A^n e^{-\lambda_A \sum X_i}} = \left(\frac{\lambda_0}{\lambda_A}\right)^n \exp\left((\lambda_A - \lambda_0) \sum X_i\right).$$

Step 3: Rejection region

We reject H_0 when Λ is small. Since $\lambda_A > \lambda_0$, the term

$$\exp\left((\lambda_A - \lambda_0) \sum X_i\right)$$

increases as $\sum X_i$ increases.

Thus:

- Λ increases as $\sum X_i$ increases
- Λ decreases as $\sum X_i$ decreases

Therefore, rejecting H_0 when Λ is small is equivalent to rejecting when

$$\sum_{i=1}^{n} X_i \text{ is small.}$$

So the rejection region has the form:

$$\boxed{RR = \left\{ \sum X_i < c \right\}}.$$

Interpretation
Recall that for an exponential distribution,

$$E[X] = \frac{1}{\lambda}.$$

Since $\lambda_A > \lambda_0$, the alternative corresponds to a *smaller mean*. Thus, under H_A, we expect smaller values of $\sum X_i$.

Intuition

So the rejection rule makes intuitive sense:

- If the observed data are unusually small, we reject H_0
- This supports the alternative that λ is larger

Distribution of the test statistic
If

$$X_1, \ldots, X_n \sim \text{Exp}(\lambda),$$

then

$$\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda).$$

Thus, the cutoff c can be chosen using the Gamma distribution so that

$$P_{H_0}\left(\sum X_i < c \right) = \alpha.$$

3 Neyman–Pearson Lemma and Most Powerful Tests

The likelihood ratio test plays a central role in hypothesis testing. The Neyman–Pearson lemma provides a theoretical justification for its optimality.

Setting. Suppose we test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta = \theta_A,$$

where both hypotheses are **simple**. Let the likelihood ratio be

$$\Lambda(x) = \frac{L(\theta_0)}{L(\theta_A)}.$$

with the Rejection Region defined by RR $= \{\Lambda < c\}$. Denote $\alpha = P_{H_0}(RR)$. Then, any other test for which the level of significance (i.e., Type I error rate) is $\leq \alpha$ has power less than or equal to that of the LRT.

Theorem

Neyman–Pearson Lemma.
Among all tests with significance level α, the likelihood ratio test that rejects H_0 when

$$\Lambda(x) < c$$

is the **most powerful (MP)** test.

Interpretation.

- The LRT maximizes power among all tests with the same Type I error rate
- It is therefore the **optimal test** for simple vs simple hypotheses

Remark

The Neyman–Pearson lemma does not guarantee:

- existence of an MP test for every α
- uniqueness of the MP test

Example

Neyman–Pearson lemma and MP tests

Remark

The Neyman–Pearson lemma guarantees a most powerful (MP) test when both the null and alternative hypotheses are **simple**. However, this situation is rare in practice, since most alternatives are composite.

Example 2: Binomial case
Suppose

$$X \sim \text{Binomial}(2, \theta),$$

and we test

$$H_0 : \theta = \frac{1}{2} \qquad \text{vs} \qquad H_A : \theta = \frac{3}{4}.$$

The likelihood is

$$P_\theta(X = x) = \binom{2}{x} \theta^x (1-\theta)^{2-x}, \quad x = 0, 1, 2.$$

Compute the likelihood ratio:

$$\frac{L(\theta_A)}{L(\theta_0)} = \frac{(3/4)^x(1/4)^{2-x}}{(1/2)^2}.$$

Evaluate for each x:

$$x = 0: \ \frac{L_A}{L_0} = \frac{1}{4}, \qquad x = 1: \ \frac{L_A}{L_0} = \frac{3}{4}, \qquad x = 2: \ \frac{L_A}{L_0} = \frac{9}{4}.$$

The likelihood ratio increases with x, so the MP test rejects for large values of X. Thus,

$$\boxed{RR = \{2\}.}$$

Type I error:

$$\alpha = P_{\theta=1/2}(X = 2) = \binom{2}{2}\left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Example 3: Normal case (known variance)
Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \quad \sigma \text{ known.}$$

We test

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_A : \mu = \mu_A.$$

The likelihood is

$$L(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right\}.$$

Consider the likelihood ratio:

$$\log\frac{L(\mu_A)}{L(\mu_0)} = -\frac{1}{2\sigma^2}\sum\left[(X_i - \mu_A)^2 - (X_i - \mu_0)^2\right].$$

After simplification, this becomes

$$\log\frac{L(\mu_A)}{L(\mu_0)} = \frac{\mu_A - \mu_0}{\sigma^2}\sum X_i + \text{constant}.$$

Thus, the likelihood ratio is a monotone function of

$$\sum X_i \quad \text{or equivalently} \quad \bar{X}.$$

Therefore:

- If $\mu_A > \mu_0$, reject for **large** $\bar{X}$
- If $\mu_A < \mu_0$, reject for **small** $\bar{X}$

So the MP test has rejection region

$$\boxed{RR = \{\bar{X} > c\} \text{ if } \mu_A > \mu_0,}$$

$$\boxed{RR = \{\bar{X} < c\} \text{ if } \mu_A < \mu_0,}$$

where c is chosen so that $P_{H_0}(\text{Reject}) = \alpha$. Equivalently, using

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

we reject when

$$\boxed{Z > z_\alpha \text{ or } Z < -z_\alpha.}$$

3.1 Uniformly Most Powerful (UMP) Tests

In practice, alternatives are usually composite, such as

$$H_A : \theta > \theta_0.$$

Definition

A test is called **uniformly most powerful (UMP)** at level α if it is most powerful for every parameter value in the alternative.

Key idea.

- A UMP test is optimal simultaneously for all θ_A in the alternative
- This is much stronger than the Neyman–Pearson result

Remark

In general:

- UMP tests often exist for **one-sided alternatives**
- UMP tests typically do **not exist** for two-sided alternatives

Example

Example 1: Exponential distribution

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda), \quad f(x;\lambda) = \lambda e^{-\lambda x},\ x > 0.$$

We test

$$H_0 : \lambda = \lambda_0 \qquad \text{vs} \qquad H_A : \lambda > \lambda_0.$$

The likelihood is

$$L(\lambda) = \lambda^n e^{-\lambda \sum X_i}.$$

Consider the likelihood ratio:

$$\frac{L(\lambda_A)}{L(\lambda_0)} = \left(\frac{\lambda_A}{\lambda_0}\right)^n \exp\big(-(\lambda_A - \lambda_0)\sum X_i\big).$$

Since $\lambda_A > \lambda_0$, the likelihood ratio is a decreasing function of $\sum X_i$.
Thus:

- smaller $\sum X_i \Rightarrow$ stronger evidence for H_A

Hence, the rejection region is

$$\boxed{RR = \left\{\sum X_i < c\right\}}$$

for some c chosen so that $P_{H_0}(\text{Reject}) = \alpha$.

Intuition

Since $E[X] = 1/\lambda$, larger λ implies smaller observations. Thus rejecting for small $\sum X_i$ is intuitive.

Example 2: Binomial (right-tailed)
Suppose

$$X \sim \text{Binomial}(n, \theta).$$

We test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta > \theta_0.$$

The likelihood is

$$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

The likelihood ratio becomes

$$\frac{L(\theta_A)}{L(\theta_0)} \propto \left(\frac{\theta_A}{\theta_0}\right)^x \left(\frac{1-\theta_A}{1-\theta_0}\right)^{n-x}.$$

Since $\theta_A > \theta_0$, this ratio increases with x. Thus:

- larger $X \Rightarrow$ stronger evidence for H_A

So the rejection region is

$$\boxed{RR = \{X > c\}}$$

for some cutoff c such that $P_{H_0}(X > c) = \alpha$.

Example 3: Binomial (left-tailed)
Suppose

$$X \sim \text{Binomial}(n, \theta),$$

and we test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta < \theta_0.$$

Using the same likelihood ratio, when $\theta_A < \theta_0$, the ratio is *decreasing* in x.
Thus:

- smaller $X \Rightarrow$ stronger evidence for H_A

So the rejection region is

$$\boxed{RR = \{X < c\}}$$

where c is chosen so that $P_{H_0}(X < c) = \alpha$.

3.2 Generalized Likelihood Ratio Test (GLRT)

The Neyman–Pearson lemma applies only to simple hypotheses. When hypotheses are composite, we use the **generalized likelihood ratio test (GLRT)**.

Setting. Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta),$$

and we test

$$H_0 : \theta \in \Theta_0 \qquad \text{vs} \qquad H_A : \theta \in \Theta_A,$$

where Θ_0 and Θ_A are disjoint sets. Let $\Omega = \Theta_0 \cup \Theta_A$.

Likelihood function.

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

Definition

The **generalized likelihood ratio** is

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Omega} L(\theta)} = \frac{L(\tilde{\theta}_0)}{L(\hat{\theta})},$$

where

$$\tilde{\theta}_0 = \arg\max_{\theta \in \Theta_0} L(\theta), \qquad \hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta).$$

Decision rule.

$$\text{Reject } H_0 \text{ if } \Lambda \leq \lambda_0,$$

where λ_0 is chosen such that

$$P_{H_0}(\Lambda \leq \lambda_0) \leq \alpha.$$

Remark

The GLRT is not necessarily uniformly most powerful (UMP).

Example

Normal mean (known variance)

Suppose

$$X_1, \ldots, X_n \sim N(\mu, \sigma_0^2), \quad \sigma_0 \text{ known}.$$

Test

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_A : \mu \neq \mu_0.$$

MLE:

$$\hat{\mu} = \bar{X}.$$

Restricted MLE:

$$\tilde{\mu}_0 = \mu_0.$$

Thus,

$$\Lambda = \frac{L(\mu_0)}{L(\bar{X})} = \exp\left(-\frac{n}{2\sigma_0^2}(\bar{X} - \mu_0)^2\right).$$

Hence,

$$-2\ln\Lambda = \frac{n(\bar{X} - \mu_0)^2}{\sigma_0^2} = Z^2,$$

where

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}.$$

So GLRT reduces to the usual two-sided z-test:

$$\boxed{\text{Reject if } |Z| > z_{\alpha/2}.}$$

Example

Shifted exponential model

Suppose

$$f(x;\theta) = e^{-(x-\theta)}\mathbf{1}\{x \geq \theta\}.$$

Test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta > \theta_0.$$

Likelihood:

$$L(\theta) = e^{-\sum(x_i - \theta)}\mathbf{1}\{\theta \leq \min x_i\}.$$

Thus, the MLE is

$$\hat{\theta} = \min X_i.$$

Under H_0, $\tilde{\theta}_0 = \theta_0$.
Hence,

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})} = \exp\big(-n(\hat{\theta} - \theta_0)\big).$$

Thus Λ is decreasing in $\min X_i$.
Therefore, the rejection region is

$$\boxed{RR = \{\min X_i > c\}}.$$

Remark

The same rejection region is obtained even if we replace

$$H_0 : \theta = \theta_0 \quad \text{by} \quad H_0 : \theta \leq \theta_0.$$

3.3 Asymptotic Distribution of GLRT

In most practical problems, the exact distribution of the likelihood ratio statistic Λ is difficult to obtain. Instead, we rely on its **asymptotic distribution**.

Setting (scalar parameter case). Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta),$$

where θ is a scalar parameter, and we want to test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta \neq \theta_0.$$

The likelihood ratio is

$$\Lambda = \frac{\prod_{i=1}^{n} f(X_i; \theta_0)}{\prod_{i=1}^{n} f(X_i; \hat{\theta}_n)},$$

where $\hat{\theta}_n$ is the MLE of θ. Note that $\Lambda = \Lambda_n$ depends on n, although we suppress this in notation.

Asymptotic assumptions. Assume:

- regularity conditions for the likelihood hold,
- the MLE satisfies
$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\big(0, I(\theta)^{-1}\big),$$
where $I(\theta)$ is the Fisher information.

Theorem

Asymptotic distribution of GLRT (Wilks' Theorem).
Under $H_0 : \theta = \theta_0$, as $n \to \infty$,

$$-2 \ln \Lambda \xrightarrow{d} \chi^2_1.$$

More generally,

$$-2 \ln \Lambda \xrightarrow{d} \chi^2_{df}, \quad df = \dim(\Omega) - \dim(\Theta_0).$$

Implication for rejection region. We use the chi-square distribution to determine the cutoff:

$$\alpha = P_{H_0}(\Lambda \leq \lambda_0) = P_{H_0}(-2 \ln \Lambda \geq -2 \ln \lambda_0).$$

Thus,

$$-2 \ln \lambda_0 = \chi^2_{df,\, 1-\alpha},$$

which implies

$$\boxed{\lambda_0 = \exp\left(-\frac{1}{2}\chi^2_{df,\,1-\alpha}\right).}$$

So the rejection rule becomes:

$$\boxed{\text{Reject } H_0 \text{ if } -2\ln\Lambda > \chi^2_{df,\,1-\alpha}.}$$

p-value.

$$\boxed{\text{p-value} \approx P\big(\chi^2_{df} > -2\ln\Lambda_{\text{obs}}\big).}$$

Intuition

The GLRT becomes practical because we can use the chi-square approximation to compute rejection regions and p-values without knowing the exact distribution of Λ.

Remark

This framework is particularly useful for **two-sided alternatives** of the form $H_A : \theta \neq \theta_0$.

3.4 Remarks: One-sided tests and signed LRT

For one-sided alternatives, we use the **signed square root LRT**:

$$r = \text{sign}(\hat{\theta}_n - \theta_0)\sqrt{-2\ln\Lambda}.$$

Under H_0,

$$r \approx N(0,1).$$

Interpretation.

- Converts the GLRT into a standard normal test
- Allows direct construction of one-sided rejection regions

p-value (one-sided).
If $\hat{\theta}_n$ lies in the direction specified by H_A, then

$$\text{one-sided p-value} \approx \frac{1}{2}P\big(\chi^2_{df} > -2\ln\Lambda_{\text{obs}}\big).$$

Remark

Thus, the one-sided p-value is approximately *half* of the two-sided p-value when the estimate is in the direction of the alternative.

Example

Example 1: GLRT for $X_1, \ldots, X_n \sim \text{IID}\, N(\mu, \sigma_0^2)$, σ_0 known

We want to test

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_A : \mu \neq \mu_0$$

at level α.

Step 1: Likelihood and MLEs

The likelihood is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(X_i - \mu)^2}{2\sigma_0^2}\right\}.$$

Under the full model, the MLE is

$$\hat{\mu} = \bar{X}.$$

Under H_0, the constrained MLE is

$$\tilde{\mu}_0 = \mu_0.$$

Step 2: Form the generalized likelihood ratio

$$\Lambda = \frac{L(\mu_0)}{L(\bar{X})}.$$

Taking logs and simplifying,

$$-2\ln\Lambda = \frac{1}{\sigma_0^2}\left[\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X})^2\right].$$

Use the identity

$$\sum_{i=1}^{n}(X_i - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$$

to obtain

$$-2\ln\Lambda = \frac{n(\bar{X} - \mu_0)^2}{\sigma_0^2}.$$

Equivalently,

$$-2\ln\Lambda = \left(\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2.$$

Step 3: Construct the rejection region

Since this is a two-sided test, we reject for large values of $-2\ln\Lambda$. Using the asymptotic null distribution,

$$-2\ln\Lambda \approx \chi_1^2.$$

So the GLRT rejection region is

$$-2\ln\Lambda > \chi^2_{1,\,1-\alpha}.$$

Substituting the formula above,

$$\left(\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2 > \chi^2_{1,\,1-\alpha}.$$

Since $\chi^2_{1,1-\alpha} = z^2_{\alpha/2}$, this is equivalent to

$$\boxed{\left|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right| > z_{\alpha/2}.}$$

Hence the GLRT is exactly the usual two-sided z-test.

Example

Example 2: GLRT for $f(x;\theta) = e^{-(x-\theta)}\mathbf{1}\{x \geq \theta\}$

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta) = e^{-(x-\theta)}\mathbf{1}\{x \geq \theta\}, \qquad \theta \in \mathbb{R}.$$

We want to test

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_A : \theta > \theta_0.$$

Step 1: Likelihood function

The likelihood is

$$L(\theta) = \prod_{i=1}^{n} e^{-(X_i-\theta)}\mathbf{1}\{X_i \geq \theta\} = e^{-\sum_{i=1}^n X_i + n\theta}\mathbf{1}\{\theta \leq X_{(1)}\},$$

where

$$X_{(1)} = \min(X_1, \ldots, X_n).$$

Step 2: Find the unrestricted MLE

For $\theta \leq X_{(1)}$,

$$L(\theta) \propto e^{n\theta},$$

which is increasing in θ. Therefore the unrestricted MLE is

$$\hat{\theta} = X_{(1)}.$$

Under H_0, the constrained value is simply

$$\tilde{\theta}_0 = \theta_0.$$

Step 3: Form the generalized likelihood ratio

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})} = \frac{e^{-\sum X_i + n\theta_0}\mathbf{1}\{\theta_0 \leq X_{(1)}\}}{e^{-\sum X_i + nX_{(1)}}}.$$

Whenever $X_{(1)} \geq \theta_0$,

$$\Lambda = e^{-n(X_{(1)}-\theta_0)}.$$

Thus Λ is a decreasing function of $X_{(1)}$.

Step 4: Construct the rejection region

Since the GLRT rejects for small values of Λ, and Λ decreases as $X_{(1)}$ increases, the rejection region must be of the form

$$\boxed{RR = \{X_{(1)} > c\}}$$

for some cutoff c chosen so that the test has level α.
So we reject H_0 when the sample minimum is sufficiently large.

Intuition

This makes intuitive sense: because every observation must satisfy $X_i \geq \theta$, increasing θ shifts the whole sample to the right. Therefore evidence for $H_A : \theta > \theta_0$ comes from a large value of the sample minimum.

Remark

The same rejection region is obtained if the null hypothesis is replaced by

$$H_0' : \theta \leq \theta_0.$$

4 Duality Between Hypothesis Testing and Confidence Intervals, Pearson's χ^2 Goodness-of-Fit Test, and Normal Q–Q Plots

4.1 Duality Between Hypothesis Testing and Confidence Intervals

Suppose we are interested in an unknown scalar parameter θ. We want to test

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_A : \theta \neq \theta_0.$$

Assume that we have an estimator $\hat{\theta}$ for θ, together with an estimated standard error $\hat{\sigma}_{\hat{\theta}}$. In large samples, suppose the standardized estimator is approximately normal:

$$\frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} \approx N(0, 1).$$

A common large-sample test is the *Wald test*. Its test statistic is

$$Z = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}}.$$

Under the null hypothesis $H_0 : \theta = \theta_0$, we then have approximately

$$Z \sim N(0, 1).$$

Rejection region for the two-sided test For a level α two-sided test, the rejection region is

$$RR = \{z : |z| > z_{\alpha/2}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ critical value of the standard normal distribution.
Thus, we reject H_0 if

$$\left| \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}} \right| > z_{\alpha/2},$$

and we fail to reject H_0 if

$$\left| \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}} \right| \leq z_{\alpha/2}.$$

Let us now carefully manipulate this inequality:

$$\left| \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}} \right| \leq z_{\alpha/2} \iff -z_{\alpha/2} \leq \frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\hat{\theta}}} \leq z_{\alpha/2}.$$

Multiplying through by $\hat{\sigma}_{\hat{\theta}} > 0$ gives

$$-z_{\alpha/2}\hat{\sigma}_{\hat{\theta}} \leq \hat{\theta} - \theta_0 \leq z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}.$$

Rearranging yields

$$\hat{\theta} - z_{\alpha/2}\hat{\sigma}_{\hat{\theta}} \leq \theta_0 \leq \hat{\theta} + z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}.$$

But, this is exactly the statement that θ_0 belongs to the interval

$$\left(\hat{\theta} - z_{\alpha/2}\hat{\sigma}_{\hat{\theta}},\, \hat{\theta} + z_{\alpha/2}\hat{\sigma}_{\hat{\theta}} \right),$$

which is the usual approximate $100(1-\alpha)\%$ confidence interval for θ.

The duality result We therefore obtain the following fundamental result.

Theorem (Duality between two-sided tests and confidence intervals). For the Wald large-sample procedure, testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_A : \theta \neq \theta_0$$

at significance level α is equivalent to checking whether θ_0 lies in the $100(1-\alpha)\%$ confidence interval for θ. More precisely:

$$\text{Fail to reject } H_0 \iff \theta_0 \in \text{CI}_{1-\alpha}(\theta),$$

and

$$\text{Reject } H_0 \iff \theta_0 \notin \text{CI}_{1-\alpha}(\theta).$$

Interpretation. A hypothesis test asks whether one particular value θ_0 is plausible. A confidence interval gives the entire set of plausible parameter values. The theorem says that these two procedures contain the same information when they are constructed from the same test statistic and the same confidence level.

Example Suppose that, based on a sample, we construct a 95% confidence interval for the population mean μ:

$$(10, 15).$$

Now consider testing

$$H_0 : \mu = 17 \qquad \text{versus} \qquad H_A : \mu \neq 17$$

at significance level $\alpha = 0.05$. Since $17 \notin (10, 15)$, the null value does not lie in the 95% confidence interval. By the duality principle, we reject H_0 at the 5% significance level. On the other hand, if we tested

$$H_0 : \mu = 12 \qquad \text{versus} \qquad H_A : \mu \neq 12,$$

then we would fail to reject H_0, because $12 \in (10, 15)$.

One-sided versions A similar relationship holds between one-sided tests and one-sided confidence intervals.

- Testing $H_0 : \theta = \theta_0$ versus $H_A : \theta > \theta_0$ corresponds to checking whether θ_0 lies in an upper confidence bound.
- Testing $H_0 : \theta = \theta_0$ versus $H_A : \theta < \theta_0$ corresponds to checking whether θ_0 lies in a lower confidence bound.

So the duality is not limited to two-sided tests. It is a general principle in inference.

4.2 Pearson's χ^2 Test and the Generalized Likelihood Ratio Test

4.2.1 Multinomial model

Suppose

$$(X_1, \ldots, X_m) \sim \text{Multinomial}(n; p_1, \ldots, p_m),$$

so that

$$P(X_1 = x_1, \ldots, X_m = x_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}, \qquad x_1 + \cdots + x_m = n.$$

Assume that prior knowledge suggests that the cell probabilities are determined by a parameter $\theta \in \mathbb{R}^k$:

$$p_j = p_j(\theta), \qquad j = 1, \ldots, m.$$

We want to test

$$H_0 : p_1 = p_1(\theta), \ldots, p_m = p_m(\theta)$$

versus

$$H_A : p_j \neq p_j(\theta) \text{ for some } j,$$

subject to the constraint

$$p_1 + \cdots + p_m = 1.$$

Under H_0, the multinomial probabilities must come from the parametric model $p_j(\theta)$. Under H_A, the probabilities are unrestricted except for summing to one.

4.2.2 Generalized likelihood ratio

The generalized likelihood ratio is

$$\Lambda = \frac{\sup_{H_0} L}{\sup_{H_A \cup H_0} L}.$$

Let us compute the numerator and denominator separately.

Likelihood under the null. Under H_0, the likelihood is

$$L_0(\theta) = \frac{n!}{x_1! \cdots x_m!} \prod_{j=1}^{m} p_j(\theta)^{x_j}.$$

Maximizing over θ gives

$$\sup_{H_0} L = \frac{n!}{x_1! \cdots x_m!} \prod_{j=1}^{m} p_j(\hat{\theta})^{x_j},$$

where $\hat{\theta}$ is the MLE under H_0.

Likelihood under the unrestricted alternative. Without the parametric restriction, the likelihood is

$$L(p_1, \ldots, p_m) = \frac{n!}{x_1! \cdots x_m!} \prod_{j=1}^{m} p_j^{x_j}, \qquad \sum_{j=1}^{m} p_j = 1.$$

Its maximum occurs at

$$\hat{p}_j = \frac{x_j}{n}, \qquad j = 1, \ldots, m,$$

the unrestricted multinomial MLEs. Therefore,

$$\sup_{H_A \cup H_0} L = \frac{n!}{x_1! \cdots x_m!} \prod_{j=1}^{m} \hat{p}_j^{x_j}.$$

Hence

$$\Lambda = \frac{\frac{n!}{x_1! \cdots x_m!} \prod_{j=1}^{m} p_j(\hat{\theta})^{x_j}}{\frac{n!}{x_1! \cdots x_m!} \prod_{j=1}^{m} \hat{p}_j^{x_j}} = \prod_{j=1}^{m} \left(\frac{p_j(\hat{\theta})}{\hat{p}_j} \right)^{x_j}.$$

4.2.3 From the LRT to Pearson's statistic

A standard asymptotic result for the generalized likelihood ratio test says that, under H_0 and for large n,

$$-2 \log \Lambda \xrightarrow{d} \chi^2_{m-1-k}.$$

Moreover, it turns out that

$$-2 \log \Lambda \approx \sum_{j=1}^{m} \frac{\{X_j - n p_j(\hat{\theta})\}^2}{n p_j(\hat{\theta})}.$$

If we define

$$O_j = X_j \qquad \text{and} \qquad E_j = n p_j(\hat{\theta}),$$

then the right-hand side becomes

$$\chi^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}.$$

This is *Pearson's chi-square statistic*. Thus, Pearson's test statistic can be viewed as a large-sample approximation to the generalized likelihood ratio test statistic in the multinomial model.

Degrees of freedom. The asymptotic reference distribution is

$$\chi^2 \overset{approx}{\sim} \chi^2_{m-1-k}.$$

The degrees of freedom are:

- m cells,
- minus 1 because the probabilities must sum to one,
- minus k because k parameters are estimated from the data.

4.2.4 Pearson's goodness-of-fit test for a general model

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta),$$

where θ is a scalar or vector parameter. We observe a sample

$$\{x_1, \ldots, x_n\},$$

and estimate the parameter by $\hat{\theta}$, for example by MLE or MOM. We now divide the sample space into m bins:

$$I_1 = (-\infty, a_1], \quad I_2 = (a_1, a_2], \quad \ldots, \quad I_m = (a_{m-1}, \infty).$$

Define

$$p_j(\theta) = P_\theta(X \in I_j) = \int_{I_j} f(x; \theta)\, dx, \qquad j = 1, \ldots, m.$$

Then, the goodness-of-fit problem becomes

$$H_0 : p_1 = p_1(\theta), \ldots, p_m = p_m(\theta)$$

versus

$$H_A : p_j \neq p_j(\theta) \text{ for some } j.$$

Under H_0, the model $f(x; \theta)$ is adequate for the observed data. Under H_A, it is not.

4.2.5 Observed and expected frequencies

For each bin I_j, define the observed frequency

$$O_j = \#\{x_i : x_i \in I_j\}.$$

The expected frequency under the fitted model is

$$E_j = nP_{\hat{\theta}}(X \in I_j) = n\, p_j(\hat{\theta}).$$

Pearson's statistic is then

$$\chi^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}.$$

Under H_0 and for large n,

$$\chi^2 \overset{approx}{\sim} \chi^2_{m-1-\dim(\theta)}.$$

Therefore, the p-value is

$$p\text{-value} = P\Big(\chi^2_{m-1-\dim(\theta)} > \chi^2_{\text{obs}}\Big).$$

If the p-value is small, we reject H_0 and conclude that the model is not adequate for the data.

Important practical remark. Pearson's statistic is particularly natural for discrete data. For continuous distributions, the result depends on how the bins are chosen, and different bin choices may lead to different conclusions. In addition, expected counts should not be too small; otherwise cells should be combined.

4.3 Detailed Examples for Pearson's χ^2 Test

Example 1: Fairness of a coin using total heads and tails

A student reported getting

$$\text{9207 heads} \qquad \text{and} \qquad \text{8743 tails}$$

in

$$\text{17,950}$$

coin tosses. We want to test whether the data are consistent with the null hypothesis that the coin is fair:

$$H_0 : p = \frac{1}{2} \qquad \text{versus} \qquad H_A : p \neq \frac{1}{2}.$$

Step 1: Observed counts. There are two categories, heads and tails:

$$O_1 = 9207, \qquad O_2 = 8743.$$

Step 2: Expected counts under H_0. If the coin is fair, then

$$P(\text{Head}) = P(\text{Tail}) = \frac{1}{2}.$$

Since the total number of tosses is $n = 17{,}950$, the expected counts are

$$E_1 = E_2 = 17{,}950 \cdot \frac{1}{2} = 8{,}975.$$

Step 3: Compute Pearson's statistic. Pearson's chi-square statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

So here,

$$\chi^2 = \frac{(9207-8975)^2}{8975} + \frac{(8743-8975)^2}{8975}.$$

Now,

$$9207 - 8975 = 232, \qquad 8743 - 8975 = -232.$$

Hence,

$$\chi^2 = \frac{232^2}{8975} + \frac{(-232)^2}{8975} = 2 \cdot \frac{53824}{8975} \approx 11.994.$$

Step 4: Degrees of freedom. There are $m = 2$ cells and no parameters are estimated under the null hypothesis, so

$$df = m - 1 = 1.$$

Step 5: Compute the p-value.

$$p\text{-value} = P(\chi_1^2 > 11.994).$$

This probability is very small, in fact less than 0.001.

Conclusion. Since the p-value is much smaller than common significance levels such as 0.05 or 0.01, we reject H_0. The observed difference between the numbers of heads and tails is too large to be explained by ordinary random variation if the coin were truly fair.

Example 2: Tossing groups of five fair coins

To save time, the student tossed groups of five coins at a time, for a total of 3590 groups, and recorded the number of heads in each group:

Number of Heads	0	1	2	3	4	5
Frequency	100	524	1080	1126	655	105

We want to test whether these data are consistent with the hypothesis that all coins are fair and independent:

$$H_0 : Y \sim \text{Binomial}(5, 1/2),$$

where Y is the number of heads in one group of five tosses.

Step 1: Compute the theoretical probabilities under H_0. If each of the five tosses is independent and fair, then

$$P(Y = j) = \binom{5}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{5-j} = \binom{5}{j} \left(\frac{1}{2}\right)^5 = \frac{\binom{5}{j}}{32}, \qquad j = 0, 1, 2, 3, 4, 5.$$

Thus,

$$(P(Y=0), P(Y=1), P(Y=2), P(Y=3), P(Y=4), P(Y=5)) = \left(\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32}\right).$$

Equivalently,

$$(P(Y=0), P(Y=1), P(Y=2), P(Y=3), P(Y=4), P(Y=5)) =$$

$$(0.03125,\ 0.15625,\ 0.3125,\ 0.3125,\ 0.15625,\ 0.03125).$$

Step 2: Compute expected counts. Since there are 3590 groups, the expected count in each category is

$$E_j = 3590 \cdot P(Y=j).$$

Therefore,

$$E_0 = 3590 \cdot \frac{1}{32} = 112.1875,$$

$$E_1 = 3590 \cdot \frac{5}{32} = 560.9375,$$

$$E_2 = 3590 \cdot \frac{10}{32} = 1121.875,$$

$$E_3 = 3590 \cdot \frac{10}{32} = 1121.875,$$

$$E_4 = 3590 \cdot \frac{5}{32} = 560.9375,$$

$$E_5 = 3590 \cdot \frac{1}{32} = 112.1875.$$

So the expected counts are

$$(E_0, E_1, E_2, E_3, E_4, E_5) = (112.1875,\ 560.9375,\ 1121.875,\ 1121.875,\ 560.9375,\ 112.1875).$$

Step 3: Compute Pearson's statistic. The observed counts are

$$(O_0, O_1, O_2, O_3, O_4, O_5) = (100, 524, 1080, 1126, 655, 105).$$

Hence,

$$\chi^2 = \sum_{j=0}^{5} \frac{(O_j - E_j)^2}{E_j}.$$

Now compute each contribution separately:

$$\frac{(100 - 112.1875)^2}{112.1875} = \frac{(-12.1875)^2}{112.1875} \approx 1.324,$$

$$\frac{(524 - 560.9375)^2}{560.9375} = \frac{(-36.9375)^2}{560.9375} \approx 2.433,$$

$$\frac{(1080-1121.875)^2}{1121.875} = \frac{(-41.875)^2}{1121.875} \approx 1.563,$$

$$\frac{(1126-1121.875)^2}{1121.875} = \frac{(4.125)^2}{1121.875} \approx 0.015,$$

$$\frac{(655-560.9375)^2}{560.9375} = \frac{(94.0625)^2}{560.9375} \approx 15.771,$$

$$\frac{(105-112.1875)^2}{112.1875} = \frac{(-7.1875)^2}{112.1875} \approx 0.460.$$

Adding all terms gives

$$\chi^2 \approx 1.324 + 2.433 + 1.563 + 0.015 + 15.771 + 0.460 = 21.566.$$

Step 4: Degrees of freedom. There are $m = 6$ categories, and no parameters are estimated from the data under the null model. Therefore,

$$df = m - 1 = 5.$$

Step 5: Compute the p-value.

$$p\text{-value} = P(\chi^2_5 > 21.566).$$

This value is very small.

Conclusion. We reject the null hypothesis that the outcomes come from independent tosses of five fair coins. The observed frequencies deviate too much from the binomial model with $n = 5$ and $p = 1/2$.

Interpretation. The largest contribution to the chi-square statistic comes from the category of 4 heads, whose observed frequency is much larger than expected. This example shows that even if the overall proportion of heads were roughly close to $1/2$, the *full distribution* of outcomes may still fail to match the binomial model.

Example 3: Horse-kick fatalities and the Poisson model

The number of fatalities from horse kicks per year was recorded for 10 army corps over 20 years, yielding 200 corps-years of data. The observed frequencies are:

# deaths	Observed (O)
0	109
1	65
2	22
3	3
4	1

Assume the number of deaths per corps-year follows a Poisson distribution:

$$Y \sim \text{Poisson}(\lambda).$$

Using maximum likelihood estimation, we are given

$$\hat{\lambda}_{MLE} = 0.61.$$

We want to test whether the Poisson model with rate $\lambda = 0.61$ is adequate for the data.

Step 1: State the hypotheses. The null hypothesis is

$$H_0 : Y \sim \text{Poisson}(0.61),$$

and the alternative is

$$H_A : \text{the distribution of } Y \text{ is not Poisson}(0.61).$$

Step 2: Compute the expected probabilities under the null model. For a Poisson random variable,

$$P(Y = j) = e^{-\lambda}\frac{\lambda^j}{j!}.$$

Using $\lambda = 0.61$:

$$P(Y = 0) = e^{-0.61}\frac{0.61^0}{0!} = e^{-0.61} \approx 0.543,$$

$$P(Y = 1) = e^{-0.61}\frac{0.61^1}{1!} \approx 0.331,$$

$$P(Y = 2) = e^{-0.61}\frac{0.61^2}{2!} \approx 0.101,$$

$$P(Y = 3) = e^{-0.61}\frac{0.61^3}{3!} \approx 0.021,$$

$$P(Y = 4) = e^{-0.61}\frac{0.61^4}{4!} \approx 0.003.$$

Step 3: Compute expected counts. Since there are 200 corps-years, multiply each probability by 200:

$$E_0 = 200(0.543) = 108.6 \approx 109,$$

$$E_1 = 200(0.331) = 66.2 \approx 66,$$

$$E_2 = 200(0.101) = 20.2 \approx 20,$$

$$E_3 = 200(0.021) = 4.2 \approx 4,$$

$$E_4 = 200(0.003) = 0.6 \approx 1.$$

Thus the expected counts are approximately

$$(E_0, E_1, E_2, E_3, E_4) = (109, 66, 20, 4, 1).$$

Step 4: Compute Pearson's statistic. Pearson's chi-square statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

So here,

$$\chi^2 = \frac{(109-109)^2}{109} + \frac{(65-66)^2}{66} + \frac{(22-20)^2}{20} + \frac{(3-4)^2}{4} + \frac{(1-1)^2}{1}.$$

Now simplify term by term:

$$\frac{(109-109)^2}{109} = 0,$$

$$\frac{(65-66)^2}{66} = \frac{1}{66} \approx 0.015,$$

$$\frac{(22-20)^2}{20} = \frac{4}{20} = 0.200,$$

$$\frac{(3-4)^2}{4} = \frac{1}{4} = 0.250,$$

$$\frac{(1-1)^2}{1} = 0.$$

Therefore,

$$\chi^2 \approx 0 + 0.015 + 0.200 + 0.250 + 0 = 0.465.$$

Step 5: Degrees of freedom. There are $m = 5$ categories, and one parameter λ was estimated from the data. Hence,

$$df = m - 1 - 1 = 3.$$

Step 6: Compute the p-value.

$$p\text{-value} = P(\chi^2_3 > 0.465).$$

This p-value is large.

Conclusion. We fail to reject H_0. The observed frequencies are very close to those predicted by a Poisson distribution with parameter $\lambda = 0.61$, so the Poisson model appears to be adequate for these data.

Interpretation. This example illustrates the main purpose of a goodness-of-fit test: we do not try to prove that a model is exactly true, but we assess whether the discrepancy between the observed data and the proposed model is too large to be explained by chance.

4.4 Normal Q–Q Plots

Another way to assess whether an assumed model is valid for the data is to visually examine a *Quantile–Quantile plot*, or *Q–Q plot*. The normal Q–Q plot is specifically designed to assess whether the data are consistent with a normal distribution. While Pearson's χ^2 test gives a formal hypothesis test, the Q–Q plot gives a visual diagnostic. It helps us understand not only whether the normal model is questionable, but also *how* it fails.

4.4.1 Construction of the normal Q–Q plot

Let the ordered sample be

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}.$$

For each i, choose a probability level

$$p_i = \frac{i - 0.5}{n}.$$

Then compute the corresponding theoretical quantile from the standard normal distribution:

$$z_i = \Phi^{-1}(p_i).$$

A normal Q–Q plot graphs the pairs

$$(z_i, X_{(i)}), \qquad i = 1, \ldots, n.$$

If the data come from a normal distribution with mean μ and standard deviation σ, then approximately

$$X_{(i)} \approx \mu + \sigma z_i,$$

so the points should lie approximately on a straight line.

4.4.2 Interpretation

1. **Straight line:** If the points lie roughly on a straight line, then the data are approximately normal.

2. **Heavier tails than normal:** If the points bend away from the line at both ends, then the data have heavier tails than a normal distribution.

3. **Lighter tails than normal:** If the points flatten at both ends compared with the line, then the data have lighter tails than a normal distribution.

4. **Right skewness:** If the upper tail rises more sharply than expected, the data are right-skewed.

5. **Left skewness:** If the lower tail falls more sharply than expected, the data are left-skewed.

Thus, the Q–Q plot is not just a yes/no normality check. It is a tool for understanding tail behavior and skewness.

4.4.3 Exercise: Interpreting Q–Q plots from simulated data

Simulate 1000 observations from each of the following distributions:

1. t_2
2. Exponential with rate 5
3. Normal$(2, 3^2)$
4. Uniform$(2, 3)$

For each sample, draw a normal Q–Q plot and discuss what you see.

Expected discussion.

- t_2**:** Since t_2 has much heavier tails than the normal distribution, the points should deviate strongly from a straight line in both tails.
- **Exponential**(5)**:** This distribution is strongly right-skewed, so the Q–Q plot should show clear curvature, especially in the upper tail.
- **Normal**$(2, 3^2)$**:** Since this is still a normal distribution, the Q–Q plot should be approximately linear. The line may have a different slope and intercept, reflecting the different mean and variance.
- **Uniform**$(2, 3)$**:** Because the uniform distribution has bounded support and lighter tails than the normal distribution, the Q–Q plot should flatten at both ends.

4.5 Summary

In this section, we developed three important ideas.

1. A two-sided level α hypothesis test is dual to a $100(1-\alpha)\%$ confidence interval: we reject the null if and only if the null value lies outside the interval.

2. Pearson's χ^2 statistic
$$\chi^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}$$
provides a large-sample goodness-of-fit test for comparing observed and expected frequencies.

3. Pearson's test can be derived from the generalized likelihood ratio test in the multinomial setting, and its asymptotic degrees of freedom are
$$m - 1 - k.$$

4. Normal Q–Q plots provide a graphical diagnostic for normality and help identify skewness, heavy tails, and light tails.

Together, these ideas show how statistical inference combines formal testing, interval estimation, and graphical diagnostics to assess uncertainty and model adequacy.

5 The Poisson Dispersion Test

This section considers whether count data are consistent with a Poisson model.

5.1 Why dispersion matters

A Poisson random variable satisfies
$$\mathbb{E}[X] = \lambda, \qquad \mathrm{Var}(X) = \lambda.$$
Thus the Poisson model has a very distinctive feature:
$$\text{mean} = \text{variance}.$$
If the sample variance is much larger than the sample mean, that suggests **overdispersion**. If it is much smaller, that suggests **underdispersion**.

5.2 General idea of the test

Suppose
$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$$
under the null. The MLE of λ is
$$\widehat{\lambda} = \bar{X}.$$
To assess fit, one compares the observed spread of the data to the spread predicted by the fitted Poisson model.

A typical dispersion statistic has the rough form
$$D = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\bar{X}},$$

possibly followed by standardization depending on the exact form of the test.

5.3 Interpretation

- If D is too large, the data are more variable than a Poisson model allows.
- If D is too small, the data are too concentrated relative to a Poisson model.

Remark

Overdispersion is extremely common in real count data because of unobserved heterogeneity, dependence, or clustering. That is why dispersion testing is a useful diagnostic step before committing to a Poisson model.

6 Probability Plots

Probability plots are graphical tools for assessing goodness of fit.

6.1 Main idea

Suppose we want to assess whether data come from a distribution with cdf F. Let the ordered sample be

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

If the model is correct, then the ordered observations should line up approximately with the theoretical quantiles

$$F^{-1}\left(\frac{k}{n+1}\right), \qquad k = 1, \ldots, n.$$

So we plot

$$X_{(k)} \quad \text{against} \quad F^{-1}\left(\frac{k}{n+1}\right).$$

6.2 What should the plot look like?

If the candidate model is appropriate, the points should lie approximately on a straight line.

6.3 Location-scale families

If the candidate distribution belongs to a location-scale family,

$$F(x) = G\left(\frac{x-\mu}{\sigma}\right),$$

then

$$X_{(k)} \approx \sigma G^{-1}\left(\frac{k}{n+1}\right) + \mu.$$

Thus the points should still be approximately linear, but the slope and intercept encode the scale and location parameters.

6.4 Why probability plots are useful

A formal goodness-of-fit test reduces all information to a single number and a single decision. A probability plot provides a richer picture:

- whether the left tail fits,
- whether the center fits,
- whether the right tail fits,
- whether the data are skewed,
- whether the tails are heavier or lighter than expected.

6.5 Interpreting common patterns

If the right tail bends upward relative to the reference line, that suggests a heavier right tail than the model predicts. An S-shaped curve often suggests skewness or mismatch in tail behavior. A few isolated points far from the trend may indicate outliers.

Intuition

Probability plots are especially powerful because they reveal *how* a model fails, not just whether a formal test rejects it.

7 Tests for Normality

Because the normal distribution plays such a central role in statistics, it is natural to test whether a sample is approximately normal.

7.1 Why normality matters

Many classical inference procedures rely on exact normality or on approximations that work best when the data are not too far from normal. Strong non-normality can affect:

- accuracy of p-values,
- reliability of confidence intervals,
- sensitivity to outliers,
- interpretation of least-squares methods.

7.2 Graphical versus formal methods

There are two major ways to assess normality:

- graphical methods, especially the normal probability plot;
- formal normality tests.

7.3 Normal probability plot

For a normal model, we compare the ordered sample values to theoretical normal quantiles:

$$X_{(k)} \quad \text{versus} \quad \Phi^{-1}\left(\frac{k}{n+1}\right).$$

If the sample is approximately normal, the points should fall close to a line.

7.4 What non-normality looks like

Typical signs include:

- curvature in both tails, suggesting heavy or light tails;
- asymmetric bending, suggesting skewness;
- isolated extreme points, suggesting outliers.

7.5 Formal normality tests

Many formal tests of normality exist. Although Chapter 9 emphasizes the general ideas more than any one specific test, the common logic is always the same:

> Measure how far the empirical behavior of the sample departs from what normality would predict.

Widely used procedures in practice include:

- Shapiro–Wilk,
- Anderson–Darling,
- Kolmogorov–Smirnov-type tests,
- tests based on skewness and kurtosis.

Remark

A normality test may reject in a very large sample because of a tiny deviation that has little practical importance. That is why the normal probability plot is often more informative than a bare p-value.

8 Concluding Remarks

Chapter 9 unifies two major themes: testing hypotheses and assessing goodness of fit.

8.1 Hypothesis testing as decision-making

A statistical test is a rule for deciding whether data are sufficiently incompatible with a null model to justify rejecting that model. The frequentist framework emphasizes:

- control of Type I error,
- analysis of Type II error and power,
- rejection regions,
- and p-values.

8.2 Likelihood as the central organizing idea

Likelihood ratios appear repeatedly:

- in the two-coin example,
- in the Neyman–Pearson lemma,
- in generalized likelihood ratio tests,
- in multinomial goodness-of-fit procedures.

This is not accidental. Likelihood is the mathematical language for asking which model makes the observed data more plausible.

8.3 Formal and graphical methods should complement each other

A central lesson of the chapter is that a single test statistic is not the whole story. Formal tests are useful, but graphical tools such as probability plots can reveal where a model fails and whether the failure is scientifically important.

8.4 Final lessons to remember

- Rejecting H_0 means the data are sufficiently incompatible with it.
- Failing to reject H_0 does not prove the null is true.
- A p-value is not the probability that the null is true.
- Confidence intervals and tests are dual constructions.
- Likelihood-ratio ideas explain the form of many classical tests.
- Goodness-of-fit assessment is strongest when formal and graphical methods are used together.

This is a personal study purposed notes, based on a lecture slide given by Prof. Ana-Maria Staicu in ST 502, NC state university and Rice (3rd ed.) Chapter 9.