One-Sample Categorical Data

Test the fairness of a die

- A casino is inspected. The undercover inspector tests the fairness of a 6-face die in \( n = 400 \) throws, denoted \( X_1, \ldots, X_n \). The dataset therefore consists of 400 digits \( X_i \in \{1, \ldots, 6\} \) with \( i = 1, \ldots, n \) indexing the throws.\(^a\)

- This is one-sample categorical data.

- A categorical variable is also called a factor and the categories are also called levels. Here, the variable die here is a factor with 6 levels \( \{1, \ldots, 6\} \).

- We test
  
  \[ H_0: \text{the die is fair} \quad \text{vs} \quad H_1: \text{the die is not fair} \]

- Assuming the throws \( X_1, \ldots, X_n \) are independent, this is the same as testing\(^b\)
  
  \[ H_0: X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(\{1, \ldots, 6\}) \]

\(^a\)Typically lower-case letters are used for denoting numbers and upper-case letters for denoting random variables. At higher levels, it is common to not make the distinct, and we will adopt this practice in these lecture notes.

\(^b\)As usual, iid = ‘independent and identically distributed’ and \( \sim \) here indicates that the variables on the LHS are distributed according to the distribution on the RHS.
Frequencies and barplots

- **Summary statistics.** The main summary statistics are the frequencies or **counts** (aka frequencies) for each of the six digits:

  \[ N_s = \#\{i : X_i = s\}, \quad s = 1, \ldots, 6 \]

  When the trials are assumed independent, these counts are jointly **sufficient**.

  Intuitively, when the trials are IID, their order does not carry information about their distribution and what is left can be characterized by \((N_1, \ldots, N_6)\).

  The counts are often organized in a **table**.

- **Graphics.** Such a table is usually plotted as either

  - a **bar chart** (aka **bar plot**)
  - a **pie chart**

  The human eye is apparently more accurate on bar charts compared to pie charts. Note that, unless the variable is **ordinal**, the order of the bars is arbitrary, so that any pattern due to the order is irrelevant.

---

The Pearson chi-squared goodness-of-fit test

- **This test** compares the **observed counts**

  \[ N_1, \ldots, N_6 \]

  with the **expected counts** under the null hypothesis

  \[ \mathbb{E}_0(N_1), \ldots, \mathbb{E}_0(N_6) \]

  (\(\mathbb{E}_0\) denotes the expectation under the null hypothesis.)

  In our example,

  \[ \mathbb{E}_0(N_1) = \cdots = \mathbb{E}_0(N_6) = \frac{n}{6} \]

- **Specifically,** the test rejects for large values of

  \[
  D = \sum_{s=1}^{6} \frac{(N_s - \mathbb{E}_0(N_s))^2}{\mathbb{E}_0(N_s)} = \sum_{s=1}^{6} \frac{(N_s - n/6)^2}{n/6}
  \]
In general, suppose we observe an i.i.d. sample from a discrete distribution \( X_1, \ldots, X_n \in \{a_1, \ldots, a_S\} \), with
\[
P(X_i = a_s) = p_s, \quad s = 1, \ldots, S
\]
The parameters \( p_1, \ldots, p_S \) define the underlying distribution and, in the context of statistical inference, are unknown.

Given a probability distribution \((p_1^0, \ldots, p_S^0)\), we want to test
\[
H_0 : (p_1, \ldots, p_S) = (p_1^0, \ldots, p_S^0)
\]
versus
\[
H_1 : (p_1, \ldots, p_S) \neq (p_1^0, \ldots, p_S^0)
\]

Define the observed counts
\[
N_s = \#\{i : X_i = a_s\}.
\]
The expected counts under the null \((H_0)\) are
\[
\mathbb{E}_0(N_s) = np_s^0.
\]
The chi-squared goodness-of-fit test rejects \(H_0\) when \(D\) below is large
\[
D = \sum_{s=1}^{S} \frac{(N_s - np_s^0)^2}{np_s^0}
\]
This test is arguably the most famous test for this hypothesis testing problem and has roots in an approximation of the likelihood ratio test. (More on this later.)

Large-sample approximation of the p-value

**Theory.** Under the null, \(D\) has asymptotically \((n \to \infty)\) the chi-squared distribution with \(S - 1\) degrees of freedom.

In the context of this class, this should be understood as a numerical approximation, meaning that
\[
\mathbb{P}_0(D \leq d) \to \mathbb{P}(\chi \leq d), \quad n \to \infty
\]
for any (fixed) \(d > 0\), where \(\chi\) has the chi-squared distribution with \(S - 1\) degrees of freedom. However, the result does not say how fast the convergence is.\(^a\)

\(^a\)Some results exist that quantify the convergence speed.
The chi-squared distribution

Let \( m \) be a positive integer, and suppose \( Z_1, \ldots, Z_m \) are i.i.d. standard normal. Then the variable

\[
Y = Z_1^2 + \cdots + Z_m^2
\]

is said to have the \textit{chi-squared distribution with} \( m \) \textit{degrees of freedom}.

It has density:

\[
f(x) = \frac{1}{2^{m/2} \Gamma(m/2)} x^{m/2 - 1} e^{-x/2}, \quad x > 0
\]

where \( \Gamma \) is the Gamma function.

Approximation of the p-value by simulation

With a computer, we do not need to rely on this so-called \textit{large-sample theory}. We can estimate the 'exact' p-value by \textit{Monte Carlo simulation}.

Let \( B \) be a large integer.

1. For \( b = 1, \ldots, B \), do the following:
   (a) Generate a sample of size \( n \), \( X_1^{(b)}, \ldots, X_n^{(b)} \), from \( (p_1^0, \ldots, p_S^0) \).
   (b) Compute
   \[
   D_b = \sum_{s=1}^S \frac{(N_s^{(b)} - np_s^0)^2}{np_s^0}, \quad \text{where} \quad N_s^{(b)} = \# \{ i : X_i^{(b)} = a_s \}
   \]

2. The estimated p-value is

\[
\frac{\# \{ b : D_b \geq D \} + 1}{B + 1}
\]

It is important to use the same sample size \( n \) when simulating the samples from the null distribution.
The (generalized) likelihood ratio test

Earlier it was mentioned that the chi-squared test is an approximation to the likelihood ratio test. The likelihood ratio test rejects for large values of $Q$

$$Q = 2 \sum_{s=1}^{S} N_s \log \left( \frac{N_s}{n \hat{p}_s} \right)$$

Note that the constant factor 2 is only there so that $Q$ also has a chi-squared distribution (under the null) in the large-sample limit.

**Theory.** Under the null, $Q$ has asymptotically $(n \to \infty)$ the chi-squared distribution with $S - 1$ degrees of freedom.

(In fact, $Q/D \to 1$ in probability under the null.)

Why prefer the chi-squared test to the likelihood ratio test? In the old days, before the advent of widely accessible computers, computing $D$ was easier than computing $Q$. In our computer age, using $D$ is not as well justified.

---

Large-sample theory vs Monte Carlo

In general, why use the large-sample theory when the null distribution can be estimated more accurately by Monte Carlo simulation?

In the old days, relying on (asymptotic) analytical approximations may have been useful. Today, a Monte Carlo simulation is often easy to implement and typically more accurate.

The course introduces techniques based on the intelligent use of the computer for performing principled statistical inference. The use of the computer has the potential to bypass complicated mathematical derivations while being more accurate.