Testing Bernoulli and Binomial parameters

2024-09-05

```
stopifnot(
  require(patchwork),
  require(glue),
  require(here),
  require(tidyverse),
  require(plotly),
  require(DT),
  require(GGally),
  require(ggforce),
  require(ggfortify)
)

tidymodels::tidymodels_prefer(quiet = TRUE)

old_theme <-theme_set(theme_minimal(base_size=9, base_family = "Helvetica"))
```

- **M1 MIDS/MFA**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

> **Objectives**

Hypothesis testing

Download 'Naissances totales par sexe' from URL `https://www.ined.fr/fichier/s_rubrique/168/t35.fr` from INED.

```
path_data <- 'DATA'
births_fr_path <- here(path_data, 't35.fr.xls')
births_fr_url <- 'https://www.ined.fr/fichier/s_rubrique/168/t35.fr.xls'

if (!file.exists(births_fr_path)) {
  download.file(births_fr_url, births_fr_path, mode = "wb")
}
```

```
births_fr <-  readxl::read_excel(births_fr_path, skip = 3)

births_fr <- births_fr[2:122, ]
```

```
births_fr <- births_fr |>
  rename(year= `Répartition par sexe et vie`,
         n_livebirths = `Ensemble des nés vivants`,
         n_live_boys = `Nés vivants - Garçons`,
         n_stillbirths = `Ensemble des enfants sans vie`,
         n_still_boys =`Enfants sans vie - Garçons`) |>
  select(year, starts_with('n_'))

#births_fr <- births_fr[1:122,]

births_fr |>
  glimpse()
```

```
Rows: 121
Columns: 5
$ year          <chr> "1901", "1902", "1903", "1904", "1905", "1906", "1907", ~
$ n_livebirths  <dbl> 917075, 904434, 884498, 877091, 865604, 864745, 829632, ~
$ n_live_boys   <dbl> 468125, 462097, 451510, 447651, 442397, 441358, 424692, ~
$ n_stillbirths <dbl> 32410, 32000, 31076, 30673, 30108, 29671, 29208, 29834, ~
$ n_still_boys  <chr> "18522", "18172", "17875", "17299", "17289", "16977", "1~
```

Null Hypothesis The probability of a live newborn baby being a boy is $p_0 = .5121244$
Alternative Hypothesis The probability of a live newborn baby being a boy is $p > p_0 = .5121244$

Data and modeling

Probability of observed data if live newborn sex is distributed according to Bernoulli(p)

If amongst n livebirths we observe n_g boys:

$$\binom{n}{n_g} p^{n_g}(1-p)^{n-n_g}$$

Compute the Likelihood Ratio for alternative $p < p_0$

$$\left(\frac{p(1-p_0)}{(1-p)p_0}\right)^{n_g} \times \left(\frac{1-p}{1-p_0}\right)^{n}$$

The Likelihood Ratio increases with respect to n_g for all values of $p > p_0$.
Comparing the likelihood ratio to a threshold amounts to compare n_g to a(nother) threshold.
Here Likelihood Ratio testing is motivated by common sense and can be justified by Theory (Neyman-Pearson's Lemma).

Definition: Error of the first kind (Type I error)

The error of the first kind occurs when the null hypothesis is true, but the test would reject it.

Definition: Error of the second kind (Type II error)

The error of the second kind occurs if the alternative hypothesis is true but the test is deciding in favor of the null-hypothesis.

The next lemma justifies our interest in Likelihood Ratio testing

Lemma (Neyman-Pearson, simplified)

When testing a *simple* null hypothesis $H_0 : X \sim P_0$ against a simple alternative $H_1 : X \sim P_1$, if there exists a threshold τ such that the test T with *critical region* $\{x : p_1(x) \geq \tau \times p_0(x)\}$ has type I error probability equal to $\alpha \in (0, 1)$, then for any test T'

$$P_0\{T'(x) = 1\} \leq \alpha \quad \Rightarrow \quad P_1\{T'(x) = 1\} \leq P_1\{T(x) = 1\}$$

Level (of significance)

The level of significance is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true. The level of significance is stated to be the probability of type I error and is preset by the researcher.

Testing a Bernoulli parameter

We think of the sex of newborns as a sequence of independent Bernoulli trials. Under the simplest model, all Bernoulli trials have the same "success" probability.In principle, we have no a priori knowledge of the "success" probability. We take p_0 as the empirical frequency of livebirth of boys throughout the century.

```
births_fr |>
  summarise(tot=sum(n_livebirths, na.rm = T),
            tot_boys=sum(n_live_boys, na.rm = T),
            msr=tot_boys/tot)
```

```
# A tibble: 1 x 3
       tot tot_boys   msr
     <dbl>    <dbl> <dbl>
1 92078262 47155839 0.512
```

We compute now for each year, the probability that a binomial random variable with size *number of livebirths* during the year and success probability p_0, exceeds the number of live-births of boys during that year, the result is denoted by `pval`.

```
p_0 <- 0.5121244

births_fr <- births_fr |>
  mutate(pval = pbinom(n_live_boys, size = n_livebirths, prob = p_0, lower.tail = F)) |>
  relocate(pval, .after = year)
```

Under our null hypothesis, `pval` is a random variable, and it is (almost) uniformly distributed over $[0, 1]$. If we want a testing procedure with type I error α, we can decide to reject the null hypothesis when `pval` (usually called the p-value) is less than α.

Agree on Type I error probability (α) equal to .05.

```
births_fr |>
  DT::datatable()  |>
  DT::formatSignif('pval', digits=3) |>
  DT::formatStyle(
  'pval',
  backgroundColor = DT::styleInterval(c(.05, 1), values = c('red', 'lightgreen', 'white'))
)
```

```
PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is instal
```

Show 10 entries Search:

	year	pval	n_livebirths	n_live_boys	n_stillbirths	n_still_boys
1	1901	0.999	917075	468125	32410	18522
2	1902	0.989	904434	462097	32000	18172
3	1903	0.999	884498	451510	31076	17875
4	1904	0.999	877091	447651	30673	17299
5	1905	0.973	865604	442397	30108	17289
6	1906	0.999	864745	441358	29671	16977
7	1907	0.656	829632	424692	29208	16675
8	1908	0.299	848982	435027	29834	16914
9	1909	0.858	824739	421882	28688	16378
10	1910	0.878	828140	423581	28566	16064

Showing 1 to 10 of 121 entries

Previous 1 2 3 4 5 … 13 Next

Throughout the 123 years in the sample we observe 25 p-values smaller than 5%. This is far more than what we expect.

Geissler data, goodness of fit testing.

From package `vcdExtra`

> Geissler (1889) published data on the distributions of boys and girls in families in Saxony, collected for the period 1876-1885. The Geissler data tabulates the family composition of 991,958 families by the number of boys and girls listed in the table supplied by Edwards

```
Geissler <- vcdExtra::Geissler
Geissler |>  glimpse()
```

```
Rows: 90
Columns: 4
$ boys  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ girls <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 0, 1, 2, 3, 4, 5, 6, 7, 8~
$ size  <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, 6, 7, 8, 9~
```

```
$ Freq  <int> 108719, 42860, 17395, 7004, 2839, 1096, 436, 161, 66, 30, 8, 3, ~
```

We isolate families of size 12.

```
big_families <- Geissler |>
  filter(size==12) |>
  select(-size, -girls)
```

There are 6115 of them.

```
big_families |>
  pivot_wider(names_from = boys, values_from = Freq) |>
  knitr::kable()
```

0	1	2	3	4	5	6	7	8	9	10	11	12
3	24	104	286	670	1033	1343	1112	829	478	181	45	7

According to our simple null hypothesis, the large families compositions should be distributed according to a binomial distribution with size 12 and success probability p_0.

We can perform a goodness of fit test for this distribution. The Chi-square goodness of fit test comes to mind

```
expected <- dbinom(0:12, 12, p_0)
observed <- big_families$Freq

chisq.test(observed, p=expected) |>
  broom::tidy() |>
  knitr::kable()
```

```
Warning in chisq.test(observed, p = expected): Chi-squared approximation may be
incorrect
```

statistic	p.value	parameter	method
130.3315	0	12	Chi-squared test for given probabilities

We merge rare events as an attempt to avoid the warning.

```
expected_collapsed <- expected[2:12]
expected_collapsed[1] <- expected_collapsed[1]+ expected[1]
expected_collapsed[11] <- expected_collapsed[11]+ expected[13]

observed_collapsed <- observed[2:12]
observed_collapsed[1] <- observed_collapsed[1]+ observed[1]
observed_collapsed[11] <- observed_collapsed[11]+ observed[13]
```

```
chisq.test(observed_collapsed, p=expected_collapsed) |>
  broom::tidy() |>
  knitr::kable()
```

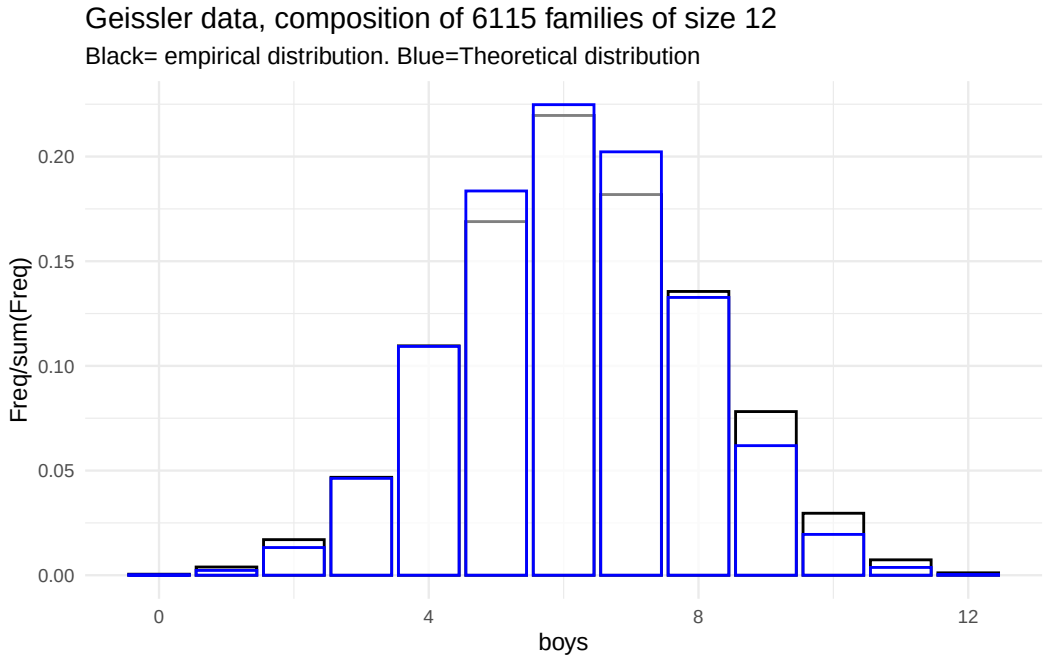
statistic	p.value	parameter	method
124.9965	0	10	Chi-squared test for given probabilities

> We are led to reject the null hypothesis for all reasonable type I error probabilities.

Let us compare the empirical distribution and the theoretical distribution of large families compisitions

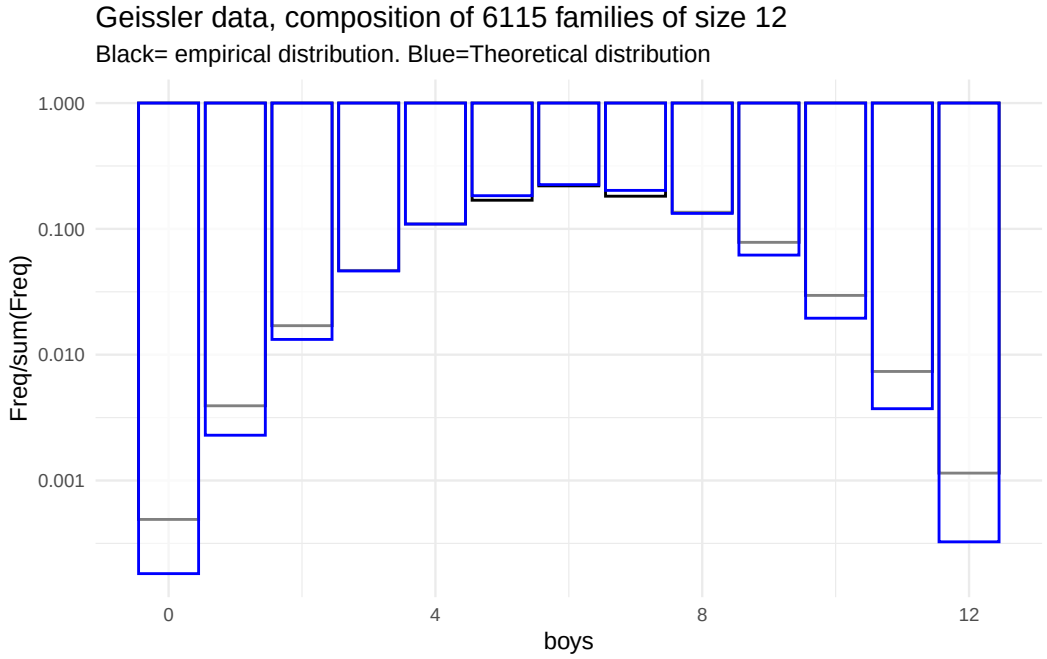
```
p <- big_families |>
  ggplot() +
  aes(x= boys, y=Freq/sum(Freq)) +
  geom_col(fill="white", color="black", alpha=.5) +
  geom_col(aes(y=expected), fill="white", color="blue", alpha=.5) +
  labs(
    title="Geissler data, composition of 6115 families of size 12",
    subtitle="Black= empirical distribution. Blue=Theoretical distribution"
  )

p
```



Using logarithmic scale on the y axis emphasizes the overdispersion of the empirical distribution.

```
p + scale_y_log10()
```


Geissler data, composition of 6115 families of size 12
Black= empirical distribution. Blue=Theoretical distribution
1.000
0.100
0.010
0.001
Freq/sum(Freq)
0
4
8
12
boys