Thomas Haslwanter

An Introduction to Statistics With Python

– Errata –

June 23, 2020

Springer

Python

p. 26, *IPython Tips #4:* Newer versions of Python no longer use Qt4, but rather Qt5. To switch between inline and external graphs, one now has to use `%matplotlib inline` and `%matplotlib qt5`.

p. 38, *Grouping* The printout of `grouped.describe()` is not formatted a bit different, and looks as follows:

```
          TV
       count      mean       std  min    25%  50%  75%  max
Gender
f        5.0  4.080000  0.769415  3.4  3.500  3.7  4.7  5.1
m        6.0  3.516667  0.926103  2.1  2.925  4.0  4.1  4.3
```

p. 42, p. 49, p. 246 *Online location of `GLM_data.zip`* This file has moved
from (old) `http://cdn.crcpress.com/downloads/C9500/GLM_data.zip`
to (new) `https://www.crcpress.com/downloads/K32369/GLM.dobson.data.zip`

Input from other formats

On p. 49, the file `data.mat` mentioned in **3.3.1 Matlab** has been added to the folder `ISP\Exercise_Solutions` in the `github` repository.

And in the corresponding code segment, the last variable should be `struct_string`, not `strunct_string`.

Background

On p. 78, the beginning of **a) Expected Value** should be

The PDF also defines the *expected value* E[X] of a continuous distribution of X:

$$E[X] = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x. \tag{0.1}$$

For discrete distributions, the integral over x is replaced by the sum over all possible values:

$$E[X] = \sum_i x_i P_i\,. \tag{0.2}$$

where x_i represents all possible values that the measured variable can have.

Distributions of One Variable

- On p. 101, only Eq. (6.11) is correct. Eq. (6.12) is wrong!
- On p. 108, the wrong figure has been inserted. **Fig. 6.11** should be

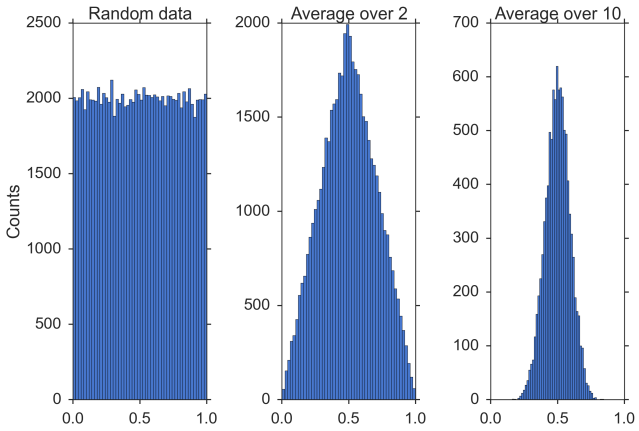

Fig. 0.1 Demonstration of the Central Limit Theorem for a uniform distribution: Left) Histogram of uniformly distributed random data between 0 and 1. Center) Histogram of average over two data points.) Right) Histogram of average over 10 data points.

- On pp. 96, 104, 116, and 118 it may be also clearer to replace *f(x)* with *p(x)*, since the functions addressed are probability distributions.
- On p. 137, care has to be taken with Fig. 7.10, explaining the ROC curve: here the blue curve (left) indicates the *patients/True positive/Population_2*. In contrast, in Figs. 7.5 and 7.6, the left curve indicates the *normals/Population_1*.

Tests of Means of Numerical Data

Comparison of Two Groups

- On p. 143, Eq.(8.1) should read

$$\begin{aligned} sd(\bar{x}_1 \pm \bar{x}_2) &= \sqrt{\mathrm{var}(\bar{x}_1) + \mathrm{var}(\bar{x}_2)} \\ &= \sqrt{\{sd(\bar{x}_1)\}^2 + \{sd(\bar{x}_2)\}^2} \\ &= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \end{aligned}$$

where $\bar{x}_i$ is the mean of the i^{th} sample, and $sd(\bar{x})$ indicates the *standard error of the mean*.

- On p. 155, the **Summary: Selecting the Right Test for Comparing Groups** should read

No. of Groups Compared	**Independent Samples**	**Paired Samples**
Groups of Nominal Data		
2 or more	Fisher's exact test or Chi-Square test	McNemar's test
Groups of Ordinal Data		
1	Wilcoxon signed rank sum test	—
2	Mann-Whitney U test	Wilcoxon signed rank sum test
3 or more	Kruskal-Wallis test	Friedman test
Groups of Continuous Data		
1	one-sample t-test or Wilcoxon signed rank sum test	—
2	Student's t-test or Mann-Whitney U test	Paired t-test or Wilcoxon signed-rank sum test
3 or more	ANOVA or Kruskal-Wallis test	Repeated Measures ANOVA or Friedman test

Table 0.1 Typical tests for statistical problems, for nominal and ordinal data. Note that the tests for comparing one group to a fixed value are the same as comparing two groups with paired samples.

Hypothetical Examples

2 groups, nominal male/female, blond-hair/black-hair. E.g. "Are females more blond than males?"

2 groups, nominal, paired 2 labs, analysis of blood samples. E.g. "Does the blood analysis from Lab1 indicate more infections than the analysis from Lab2?"

1 group, ordinal Sequence of giant-planets. E.g. "In our solar system, are giant planets further out than average in the sequence of planets?"

2 groups, ordinal Jamaican/American, ranking 100m sprint. E.g. "Are Jamaican sprinters more successful than American sprinters?"

2 groups, ordinal, paired sprinters, before/after diet. E.g. "Does a chocolate diet make sprinters more successful?"

3 groups, ordinal single/married/divorces, ranking 100m sprint. E.g. "Does the marital status have an effect on the success of sprinters?"

3 groups, ordinal, paired sprinters, before/after diet. E.g. "Does a rice diet make Chinese sprinters more successful?"

1 group, continuous Average calory intake. E.g. "Do our children eat more than they should?"

2 groups, continuous male/female, IQ. E.g. "Are women more intelligent than men?"

2 groups, continuous, paired male, looking at sport cars. E.g. "Does looking at sports cars raise the male heart-beat?

3 groups, continuous Tyrolians, Viennese, Styrians; IQ. E.g. "Are Tyrolians smarter than people from other Austrian federal states?"

3 groups, continuous, paired Tyrolians, Viennese, Styrians; looking at mountains. E.g. "Does looking at mountains raise the heartbeat of Tyrolians more than those of other people?"

2-factor ANOVA male/female, looking at diamonds. E.g. "Does looking at diamonds raise the female heart-beat more than the male?

- On p. 157, The heading of the first exercise in 8.1 should be
  One sample t-test for the mean and **Wilcoxon signed rank sum test**

Tests on Categorical Data

- On p. 170, in Table 9.5, *Subject 9* should be listed only once.
- On p. 171, the code line `obs = [[a,b], [c, d]]` is wrong. However, the ISP-Quantlet listed there is correct.

Linear Regression Models

On p. 191, Eqs. (11.18) and (11.19) have left away the offset, and should read

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \beta_0 + \mathbf{x}_i^{\mathrm{T}}\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

and

$$\mathbf{y} = \beta_0 + \mathbf{X}\beta + \varepsilon,$$

And on p. 192, in the last paragraph of Section a), the *simple linear regression* is described by

"$p = 1$".

and not by

"$i = 1$ and $p = 1$"

Solutions: 6.3 Other Continuous Distributions

On p. 253, the calculation of the F-value is incorrect. The correct calculation should read

```
fval = np.var(apples1, ddof=1)/np.var(apples2, ddof=1)
```