

3 Univariate Statistics

3.1 Introduction

3.2 Empirical Distributions

The standard deviation describes the typical spread of the observations around the mean.

Although the variance has the disadvantage of not having the same dimensions as the original data, it is used in many applications instead of the standard deviation.

3.3 Examples of Empirical Distributions

In practice, the square root of the total number of observations `len(corg)` (rounded to the nearest integer using `round`) is often used as the number of bins.

As an alternative way of plotting the data, the empirical cumulative distribution can be displayed using the argument `cumulative=True` (Fig. 3.1 b):

The third measure in this context is the mode, which is approximately the midpoint of the interval with the highest frequency.

```
plt.hist(sodium,bins=11)
```

3.4 Theoretical Distributions

The probability distribution density $f(x)$ defines the probability that the variable has a value in the vicinity of x .

3.6 Hypothesis Testing

The *p-value* of a hypothesis test is the probability (under the null hypothesis) of observing values at least as extreme as observed for the test statistics (Fig. 3.14).

3.7 The t-Test

This can be accomplished using `t.ppf()`, which yields the inverse of the *t*-distribution function with `na+nb-2` degrees of freedom at a 5% significance level.

The function `t.ppf()` yields the inverse of the t -distribution function with $na+nb-2$ degrees of freedom at a 5% significance level.

3.8 The F-Test

3.9 The χ^2 -Test

3.10 The Kolmogorov–Smirnov Test

The output variable `kscalc=0.0757` corresponds to `kscalc` in our experiment without using `kstest`. The output variable `kscrit=0.1723` differs slightly from that in Tab. 3.1 since `kstest` uses a slightly more precise approximation for the critical value for sample sizes larger than 40 from Miller (1956).

3.11 The Mann–Whitney Test

3.12 The Ansari–Bradley Test

3.13 Distribution Fitting

The object `gm` contains several **attributes**, including the mean `gm.means_` and the covariances `gm.covariances_`, which we use to calculate the probability density function y of the mixed distribution:

The object `gm` also contains information on the relative mixing proportions of the two distributions in the **attribute** `gm.weights_`.

3.14 Error Analysis

```
rng = np.random.default_rng(100)

[1.66367553  3.8346337  4.5712811  4.21596047
 1.95792604  5.006513   4.45218349  4.45746018
 4.5175939   5.05652086  6.76445859  2.48276032]

4.081747265344059
1.4366573969508232

0.4147272674314124
```

The frequency distribution has a mean of 4.1 grams per ton and a standard deviation of 1.4 grams per ton. This wide range includes the true value of 3.4 grams per ton. If I were to take 12 repeated measurements of the gold

content, the means of these measurements would have a standard deviation of approximately 0.4 grams per ton, which is the standard error of the mean.

As expected, the value is much smaller than the corresponding value for $n=12$.

Recommended Reading

Figure captions