# Statistical Inference

# 4

## 4.1 Introduction

There is not only one way to address the problem of statistical inference. In fact, there are two main approaches to statistical inference: the frequentist and Bayesian approaches. Their differences are subtle but fundamental:

- In the case of the *frequentist approach*, the main assumption is that there is a population, which can be represented by several parameters, from which we can obtain numerous random samples. Population parameters are fixed but they are not accessible to the observer. The only way to derive information about these parameters is to take a sample of the population, to compute the parameters of the sample, and to use statistical inference techniques to make probable propositions regarding population parameters.
- The *Bayesian approach* is based on a consideration that data are fixed, not the result of a repeatable sampling process, but parameters describing data can be described probabilistically. To this end, Bayesian inference methods focus on producing parameter distributions that represent all the knowledge we can extract from the sample and from prior information about the problem.

A deep understanding of the differences between these approaches is far beyond the scope of this chapter, but there are many interesting references that will enable you to learn about it [1]. What is really important is to realize that the approaches are based on different assumptions which determine the validity of their inferences. The assumptions are related in the first case to a sampling process; and to a statistical model in the second case. Correct inference requires these assumptions to be correct. The fulfillment of this requirement is not part of the method, but it is the responsibility of the data scientist.

In this chapter, to keep things simple, we will only deal with the first approach, but we suggest the reader also explores the second approach as it is well worth it!



L. Igual and S. Seguí, *Introduction to Data Science*,
Undergraduate Topics in Computer Science, DOI 10.1007/978-3-319-50017-1_4

## 4.2 Statistical Inference: The Frequentist Approach

As we have said, the ultimate objective of statistical inference, if we adopt the frequentist approach, is to produce probable propositions concerning population parameters from analysis of a sample. The most important classes of propositions are as follows:

- Propositions about *point estimates*. A point estimate is a particular value that best approximates some parameter of interest. For example, the mean or the variance of the sample.
- Propositions about *confidence intervals* or *set estimates*. A confidence interval is a range of values that best represents some parameter of interest.
- Propositions about the acceptance or rejection of a *hypothesis*.

In all these cases, the production of propositions is based on a simple assumption: we can estimate the probability that the result represented by the proposition has been caused by chance. The estimation of this probability by sound methods is one of the main topics of statistics.

The development of traditional statistics was limited by the scarcity of computational resources. In fact, the only computational resources were mechanical devices and human computers, teams of people devoted to undertaking long and tedious calculations. Given these conditions, the main results of classical statistics are theoretical approximations, based on idealized models and assumptions, to measure the effect of chance on the statistic of interest. Thus, concepts such as the *Central Limit Theorem*, the *empirical sample distribution* or the *t-test* are central to understanding this approach.

The development of modern computers has opened an alternative strategy for measuring chance that is based on simulation; producing computationally intensive methods including resampling methods (such as bootstrapping), Markov chain Monte Carlo methods, etc. The most interesting characteristic of these methods is that they allow us to treat more realistic models.

## 4.3 Measuring the Variability in Estimates

Estimates produced by descriptive statistics are not equal to the *truth* but they are better as more data become available. So, it makes sense to use them as central elements of our propositions and to measure its variability with respect to the sample size.

### 4.3.1 Point Estimates

Let us consider a dataset of accidents in Barcelona in 2013. This dataset can be downloaded from the OpenDataBCN website,[1] Barcelona City Hall's open data service. Each register in the dataset represents an accident via a series of features: weekday, hour, address, number of dead and injured people, etc. This dataset will represent our population: the set of all reported traffic accidents in Barcelona during 2013.

#### 4.3.1.1 Sampling Distribution of Point Estimates

Let us suppose that we are interested in describing the daily number of traffic accidents in the streets of Barcelona in 2013. If we have access to the *population*, the computation of this parameter is a simple operation: the total number of accidents divided by 365.

In [1]:
```
data = pd.read_csv("files/ch04/ACCIDENTS_GU_BCN_2013.csv")
data['Date'] = data[u'Dia de mes'].apply(lambda x: str(x))
               + '-' +
               data[u'Mes de any'].apply(lambda x: str(x))
data['Date'] = pd.to_datetime(data['Date'])
accidents = data.groupby(['Date']).size()
print accidents.mean()
```

Out[1]:
```
Mean: 25.9095
```

But now, for illustrative purposes, let us suppose that we only have access to a limited part of the data (the *sample*): the number of accidents during *some* days of 2013. Can we still give an approximation of the population mean?

The most intuitive way to go about providing such a mean is simply to take the *sample mean*. The sample mean is a point estimate of the population mean. If we can only choose one value to estimate the population mean, then this is our best guess.

The problem we face is that estimates generally vary from one sample to another, and this sampling variation suggests our estimate may be close, but it will not be exactly equal to our parameter of interest. How can we measure this variability?

In our example, because we have access to the population, we can empirically build the *sampling distribution of the sample mean*[2] for a given number of observations. Then, we can use the sampling distribution to compute a measure of the variability.

In Fig. 4.1, we can see the empirical sample distribution of the mean for $s = 10.000$ samples with $n = 200$ observations from our dataset. This empirical distribution has been built in the following way:

[1] http://opendata.bcn.cat/.

[2] Suppose that we draw all possible samples of a given size from a given population. Suppose further that we compute the mean for each sample. The probability distribution of this statistic is called the *mean sampling distribution*.

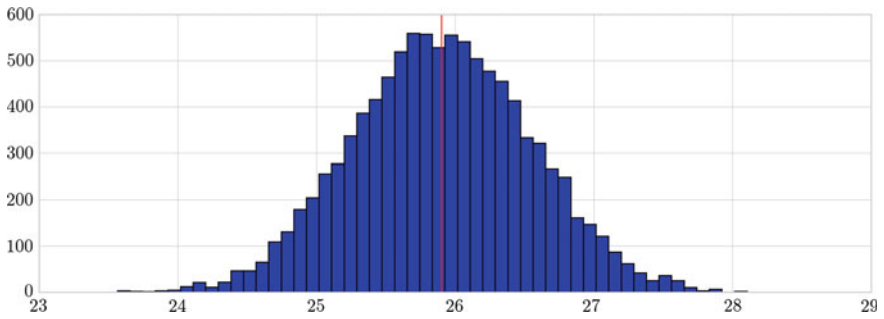


**Fig. 4.1** Empirical distribution of the sample mean. In *red*, the mean value of this distribution

1. Draw $s$ (a large number) independent samples $\{\mathbf{x}^1, \ldots, \mathbf{x}^s\}$ from the population where each element $\mathbf{x}^j$ is composed of $\{x_i^j\}_{i=1,\ldots,n}$.
2. Evaluate the sample mean $\hat{\mu}^j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$ of each sample.
3. Estimate the sampling distribution of $\hat{\mu}$ by the empirical distribution of the sample replications.

In [2]:

```
# population
df = accidents.to_frame()
N_test = 10000
elements = 200
# mean array of samples
means = [0] * N_test
# sample generation
for i in range(N_test):
    rows = np.random.choice(df.index.values, elements)
    sampled_df = df.ix[rows]
    means[i] = sampled_df.mean()
```

In general, given a point estimate from a sample of size $n$, we define its *sampling distribution* as the distribution of the point estimate based on samples of size $n$ from its population. This definition is valid for point estimates of other population parameters, such as the population median or population standard deviation, but we will focus on the analysis of the sample mean.

The sampling distribution of an estimate plays an important role in understanding the real meaning of propositions concerning point estimates. It is very useful to think of a particular point estimate as being drawn from such a distribution.

#### 4.3.1.2 The Traditional Approach

In real problems, we do not have access to the real population and so estimation of the sampling distribution of the estimate from the empirical distribution of the sample replications is not an option. But this problem can be solved by making use of some theoretical results from traditional statistics.

It can be mathematically shown that given $n$ independent observations $\{x_i\}_{i=1,..,n}$ of a population with a standard deviation $\sigma_x$, the standard deviation of the sample mean $\sigma_{\bar{x}}$, or *standard error*, can be approximated by this formula:

$$SE = \frac{\sigma_x}{\sqrt{n}}$$

The demonstration of this result is based on the Central Limit Theorem: an old theorem with a history that starts in 1810 when Laplace released his first paper on it.

This formula uses the standard deviation of the population $\sigma_x$, which is not known, but it can be shown that if it is substituted by its empirical estimate $\hat{\sigma}_x$, the estimation is sufficiently good if $n > 30$ and the population distribution is not skewed. This allows us to estimate the standard error of the sample mean even if we do not have access to the population.

So, how can we give a measure of the variability of the sample mean? The answer is simple: by giving the **empirical standard error of the mean distribution**.

In [3]:
```
rows = np.random.choice(df.index.values, 200)
sampled_df = df.ix[rows]
est_sigma_mean = sampled_df.std()/math.sqrt(200)

print 'Direct estimation of SE from one sample of
      200 elements:', est_sigma_mean[0]
print 'Estimation of the SE by simulating 10000 samples of
      200 elements:', np.array(means).std()
```

Out[3]:
```
Direct estimation of SE from one sample of 200 elements: 0.6536
Estimation of the SE by simulating 10000 samples of 200
elements: 0.6362
```

Unlike the case of the sample mean, there is no formula for the standard error of other interesting sample estimates, such as the median.

#### 4.3.1.3 The Computationally Intensive Approach

Let us consider from now that our full dataset is a sample from a hypothetical population (this is the most common situation when analyzing real data!).

A modern alternative to the traditional approach to statistical inference is the bootstrapping method [2]. In the bootstrap, we draw $n$ observations *with replacement* from the original data to create a bootstrap sample or resample. Then, we can calculate the mean for this resample. By repeating this process a large number of times, we can build a good approximation of the mean sampling distribution (see Fig. 4.2).

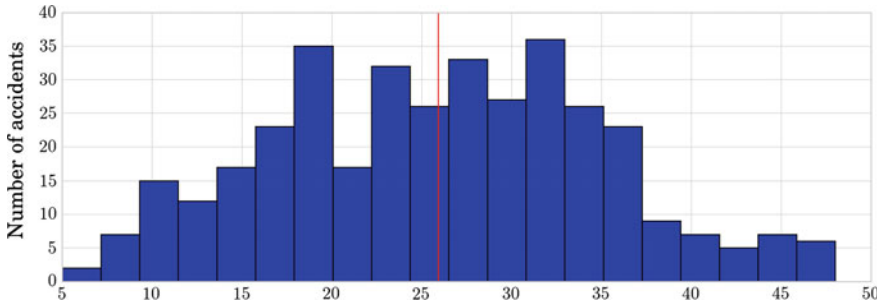


**Fig. 4.2** Mean sampling distribution by bootstrapping. In *red*, the mean value of this distribution

In [4]:
```
def meanBootstrap(X, numberb):
    x = [0]*numberb
    for i in range(numberb):
        sample = [X[j]
                  for j
                  in np.random.randint(len(X), size=len(X))
                 ]
        x[i] = np.mean(sample)
    return x
m = meanBootstrap(accidents, 10000)
print "Mean estimate:", np.mean(m)
```

Out[4]:
```
Mean estimate: 25.9094
```

The basic idea of the bootstrapping method is that the observed sample contains sufficient information about the underlying distribution. So, the information we can extract from resampling the sample is a good approximation of what can be expected from resampling the population.

The bootstrapping method can be applied to other simple estimates such as the median or the variance and also to more complex operations such as estimates of censored data.[3]

### 4.3.2 Confidence Intervals

A point estimate $\Theta$, such as the sample mean, provides a *single plausible value for a parameter*. However, as we have seen, a point estimate is rarely perfect; usually there is some error in the estimate. That is why we have suggested using the standard error as a measure of its variability.

Instead of that, a next logical step would be to provide *a plausible range of values* for the parameter. A plausible range of values for the sample parameter is called a *confidence interval*.

[3]Censoring is a condition in which the value of observation is only partially known.

We will base the definition of *confidence interval* on two ideas:

1. Our point estimate is the most plausible value of the parameter, so it makes sense to build the confidence interval around the point estimate.
2. The *plausibility* of a range of values can be defined from the sampling distribution of the estimate.

For the case of the mean, the Central Limit Theorem states that its sampling distribution is normal:

**Theorem 4.1** *Given a population with a finite mean $\mu$ and a finite non-zero variance $\sigma^2$, the sampling distribution of the mean approaches a normal distribution with a mean of $\mu$ and a variance of $\sigma^2/n$ as n, the sample size, increases.*

In this case, and in order to define an interval, we can make use of a well-known result from probability that applies to normal distributions: roughly 95% of the time our estimate will be within 1.96 standard errors of the true mean of the distribution. If the interval spreads out 1.96 standard errors from a normally distributed point estimate, intuitively we can say that we are *roughly 95% confident that we have captured the true parameter*.

$$CI = [\Theta - 1.96 \times SE, \Theta + 1.96 \times SE]$$

In [5]:
```
m = accidents.mean()
se = accidents.std()/math.sqrt(len(accidents))
ci = [m - se*1.96, m + se*1.96]
print "Confidence interval:", ci
```

Out[5]:
```
Confidence interval: [24.975, 26.8440]
```

Suppose we want to consider confidence intervals where the confidence level is somewhat higher than 95%: perhaps we would like a confidence level of 99%. To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58 (it can be shown that 99% of the time a normal random variable will be within 2.58 standard deviations of the mean).

In general, if the point estimate follows the normal model with standard error $SE$, then a confidence interval for the population parameter is

$$\Theta \pm z \times SE$$

where $z$ corresponds to the confidence level selected:

| Confidence Level | 90% | 95% | 99% | 99.9% |
|---|---|---|---|---|
| $z$ Value | 1.65 | 1.96 | 2.58 | 3.291 |

This is how we would compute a 95% confidence interval of the sample mean using bootstrapping:

1. Repeat the following steps for a large number, $s$, of times:

   a. Draw $n$ observations with replacement from the original data to create a bootstrap sample or resample.
   b. Calculate the mean for the resample.

2. Calculate the mean of your $s$ values of the sample statistic. This process gives you a "bootstrapped" estimate of the sample statistic.
3. Calculate the standard deviation of your $s$ values of the sample statistic. This process gives you a "bootstrapped" estimate of the SE of the sample statistic.
4. Obtain the 2.5th and 97.5th percentiles of your $s$ values of the sample statistic.

In [6]:
```
m = meanBootstrap(accidents, 10000)
sample_mean = np.mean(m)
sample_se =  np.std(m)

print "Mean estimate:", sample_mean
print "SE of the estimate:",  sample_se

ci = [np.percentile(m, 2.5), np.percentile(m, 97.5)]
print  "Confidence interval:", ci
```

Out[6]:
```
Mean estimate: 25.9039
SE of the estimate: 0.4705
Confidence interval: [24.9834, 26.8219]
```

#### 4.3.2.1 But What Does "95% Confident" Mean?

The real meaning of "confidence" is not evident and it must be understood from the point of view of the generating process.

Suppose we took many (infinite) samples from a population and built a 95% confidence interval from each sample. Then about 95% of those intervals would contain the actual parameter. In Fig. 4.3 we show how many confidence intervals computed from 100 different samples of 100 elements from our dataset contain the real population mean. If this simulation could be done with infinite different samples, 5% of those intervals would not contain the true mean.

So, when faced with a sample, the correct interpretation of a confidence interval is as follows:

> In 95% of the cases, when I compute the 95% confidence interval from this sample, the true mean of the population will fall within the interval defined by these bounds: $\pm 1.96 \times SE$.

We cannot say either that our specific sample contains the true parameter or that the interval has a 95% chance of containing the true parameter. That interpretation would not be correct under the assumptions of traditional statistics.

## 4.4 Hypothesis Testing

Giving a measure of the variability of our estimates is one way of producing a statistical proposition about the population, but not the only one. R.A. Fisher (1890–1962) proposed an alternative, known as *hypothesis testing*, that is based on the concept of *statistical significance*.

Let us suppose that a deeper analysis of traffic accidents in Barcelona results in a difference between 2010 and 2013. Of course, the difference could be caused only by chance, because of the variability of both estimates. But it could also be the case that traffic conditions were very different in Barcelona during the two periods and, because of that, data from the two periods can be considered as belonging to two different populations. Then, the relevant question is: Are the observed effects real or not?

Technically, the question is usually translated to: *Were the observed effects statistically significant*?

The process of determining the statistical significance of an effect is called *hypothesis testing*.

This process starts by simplifying the options into two competing hypotheses:

- $H_0$: The mean number of daily traffic accidents is the same in 2010 and 2013 (there is only one population, one true mean, and 2010 and 2013 are just different samples from the same population).
- $H_A$: The mean number of daily traffic accidents in 2010 and 2013 is different (2010 and 2013 are two samples from two different populations).

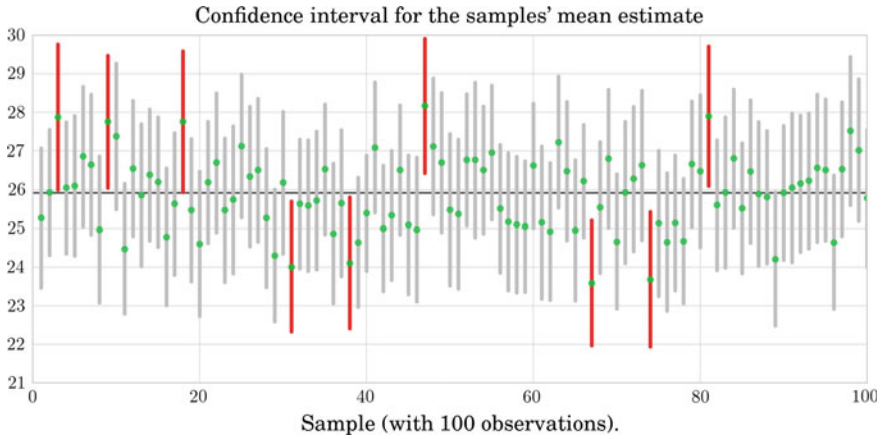


**Fig. 4.3** This graph shows 100 sample means (*green points*) and its corresponding confidence intervals, computed from 100 different samples of 100 elements from our dataset. It can be observed that a few of them (those in *red*) do not contain the mean of the population (*black horizontal line*)

We call $H_0$ the *null hypothesis* and it represents a *skeptical* point of view: the effect we have observed is due to chance (due to the specific sample bias). $H_A$ is the *alternative hypothesis* and it represents the other point of view: the effect is real.

The general rule of frequentist hypothesis testing: we will not *discard* $H_0$ (and hence we will not consider $H_A$) unless the observed effect is *implausible* under $H_0$.

### 4.4.1 Testing Hypotheses Using Confidence Intervals

We can use the concept represented by *confidence intervals* to measure the *plausibility* of a hypothesis.

We can illustrate the evaluation of the hypothesis setup by comparing the mean rate of traffic accidents in Barcelona during 2010 and 2013:

In [7]:
```
data = pd.read_csv("files/ch04/ACCIDENTS_GU_BCN_2010.csv",
                   encoding='latin-1')

# Create a new column which is the date
data['Date'] = data['Dia de mes'].apply(lambda x: str(x))
               + '-' +
               data['Mes de any'].apply(lambda x: str(x))
data2 = data['Date']
counts2010 = data['Date'].value_counts()
print '2010: Mean', counts2010.mean()

data = pd.read_csv("files/ch04/ACCIDENTS_GU_BCN_2013.csv",
                   encoding='latin-1')

# Create a new column which is the date
data['Date'] = data['Dia de mes'].apply(lambda x: str(x))
               + '-' +
               data['Mes de any'].apply(lambda x: str(x))
data2 = data['Date']
counts2013 = data['Date'].value_counts()
print '2013: Mean', counts2013.mean()
```

Out[7]:
```
2010: Mean 24.8109
2013: Mean 25.9095
```

This estimate suggests that in 2013 the mean rate of traffic accidents in Barcelona was higher than it was in 2010. But is this effect statistically significant?

Based on our sample, the 95% confidence interval for the mean rate of traffic accidents in Barcelona during 2013 can be calculated as follows:

In [8]:
```
n = len(counts2013)
mean = counts2013.mean()
s = counts2013.std()
ci = [mean - s*1.96/np.sqrt(n),  mean + s*1.96/np.sqrt(n)]
print '2010 accident rate estimate:', counts2010.mean()
print '2013 accident rate estimate:', counts2013.mean()
print 'CI for 2013:',ci
```

```
Out[8]: 2010 accident rate estimate: 24.8109
        2013 accident rate estimate: 25.9095
        CI for 2013: [24.9751, 26.8440]
```

Because the 2010 accident rate estimate does not fall in the range of plausible values of 2013, we say the alternative hypothesis cannot be discarded. That is, it cannot be ruled out that in 2013 the mean rate of traffic accidents in Barcelona was higher than in 2010.

Interpreting CI Tests

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject $H_0$ unless we have strong evidence against it. But what precisely does strong evidence mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject $H_0$ more than 5% of the time. This corresponds to a *significance level* of $\alpha = 0.05$. In this case, the correct interpretation of our test is as follows:

> If we use a 95% confidence interval to test a problem where the null hypothesis is true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter. This happens about 5% of the time (2.5% in each tail).

### 4.4.2 Testing Hypotheses Using $p$-Values

A more advanced notion of *statistical significance* was developed by R.A. Fisher in the 1920s when he was looking for a test to decide whether variation in crop yields was due to some specific intervention or merely random factors beyond experimental control.

Fisher first assumed that fertilizer caused no difference (*null hypothesis*) and then calculated $P$, the probability that an observed yield in a fertilized field would occur if fertilizer had no real effect. This probability is called the *p-value*.

The $p$-value is the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis is true. We typically use a summary statistic of the data to help compute the $p$-value and evaluate the hypotheses.

Usually, if $P$ is less than 0.05 (the chance of a fluke is less than 5%) the result is declared *statistically significant*.

It must be pointed out that this choice is rather arbitrary and should not be taken as a scientific truth.

The goal of classical hypothesis testing is to answer the question, "*Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?*" Here is how we answer that question:

- The first step is to quantify the size of the apparent effect by choosing a test statistic. In our case, the apparent effect is a difference in accident rates, so a natural choice for the test statistic is the **difference in means between the two periods**.

- The second step is to define a *null hypothesis*, which is a model of the system based on the assumption that the apparent effect is not real. In our case, the null hypothesis is that there is no difference between the two periods.
- The third step is to compute a *p-value*, which is the probability of seeing the apparent effect if the null hypothesis is true. In our case, we would compute the difference in means, then compute the probability of seeing a difference as big, or bigger, under the null hypothesis.
- The last step is to *interpret the result*. If the $p$-value is low, the effect is said to be *statistically significant*, which means that it is unlikely to have occurred by chance. In this case we infer that the effect is more likely to appear in the larger population.

In our case, the test statistic can be easily computed:

In [9]:
```
m= len(counts2010)
n= len(counts2013)
p = (counts2013.mean() - counts2010.mean())
print 'm:', m, 'n:', n
print 'mean difference: ', p
```

Out[9]:
```
m: 365 n: 365
mean difference: 1.0986
```

To approximate the $p$-value , we can follow the following procedure:

1. Pool the distributions, generate samples with size n and compute the difference in the mean.
2. Generate samples with size n and compute the difference in the mean.
3. Count how many differences are larger than the observed one.

In [10]:
```
# pooling distributions
x = counts2010
y = counts2013
pool = np.concatenate([x, y])
np.random.shuffle(pool)

#sample generation
import random
N = 10000 # number of samples
diff = range(N)
for i in range(N):
    p1 = [random.choice(pool) for _ in xrange(n)]
    p2 = [random.choice(pool) for _ in xrange(n)]
    diff[i] = (np.mean(p1) - np.mean(p2))
```

In [11]:
```
# counting differences larger than the observed one
diff2 = np.array(diff)
w1 = np.where(diff2 > p)[0]

print 'p-value (Simulation)=', len(w1)/float(N),
      '(', len(w1)/float(N)*100 ,'%)', 'Difference =', p
if (len(w1)/float(N)) < 0.05:
    print 'The effect is likely'
else:
    print 'The effect is not likely'
```

Out[11]:
```
p-value (Simulation)= 0.0485 ( 4.85%) Difference = 1.098
The effect is likely
```

Interpreting $P$-Values

A $p$-value is the probability of an observed (or more extreme) result arising only from chance.

If $P$ is less than 0.05, there are two possible conclusions: there is a real effect or the result is an improbable fluke. *Fisher's method offers no way of knowing which is the case*.

We must not confuse the odds of getting a result (if a hypothesis is true) with the odds of favoring the hypothesis if you observe that result. If $P$ is less than 0.05, we cannot say that this means that it is 95% certain that the observed effect is real and could not have arisen by chance. Given an observation $E$ and a hypothesis $H$, $P(E|H)$ and $P(H|E)$ are not the same!

Another common error equates *statistical significance* to *practical importance/ relevance*. When working with large datasets, we can detect statistical significance for small effects that are meaningless in practical terms.

We have defined the effect as *a difference in mean as large or larger than* $\delta$, *considering the sign*. A test like this is called *one sided*.

If the relevant question is whether *accident rates are different*, then it makes sense to test the absolute difference in means. This kind of test is called two sided because it counts both sides of the distribution of differences.

Direct Approach

The formula for the standard error of the absolute difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean can be approximated by:

$$SE_{\bar{x}_1} = \frac{\sigma_1}{\sqrt{n_1}}$$

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This would allow us to define a direct test with the 95% confidence interval.

## 4.5 But Is the Effect $E$ Real?

We do not yet have an answer for this question! We have defined a null hypothesis $H_0$ (the effect is not real) and we have computed the probability of the observed effect under the null hypothesis, which is $P(E|H_0)$, where $E$ is an effect as big as or bigger than the apparent effect and a $p$-value .

We have stated that from the frequentist point of view, we cannot consider $H_A$ unless $P(E|H_0)$ is less than an arbitrary value. But the real answer to this question must be based on comparing $P(H_0|E)$ to $P(H_A|E)$, not on $P(E|H_0)$! One possible solution to these problems is to use *Bayesian reasoning*; an alternative to the frequentist approach.

No matter how many data you have, you will still depend on intuition to decide how to interpret, explain, and use that data. Data cannot speak by themselves. Data scientists are interpreters, offering one interpretation of what the useful narrative story derived from the data is, if there is one at all.

## 4.6 Conclusions

In this chapter we have seen how we can approach the problem of making probable propositions regarding population parameters.

We have learned that in some cases, there are theoretical results that allow us to compute a measure of the variability of our estimates. We have called this approach the "traditional approach". Within this framework, we have seen that the sampling distribution of our parameter of interest is the most important concept when understanding the real meaning of propositions concerning parameters.

We have also learned that the traditional approach is not the only alternative. The "computationally intensive approach", based on the bootstrap method, is a relatively new approach that, based on intensive computer simulations, is capable of computing a measure of the variability of our estimates by applying a resampling method to our data sample. Bootstrapping can be used for computing variability of almost any function of our data, with its only downside being the need for greater computational resources.

We have seen that propositions about parameters can be classified into three classes: propositions about point estimates, propositions about set estimates, and propositions about the acceptance or the rejection of a hypothesis. All these classes are related; but today, set estimates and hypothesis testing are the most preferred.

Finally, we have shown that the production of probable propositions is not error free, even in the presence of big data. For these reason, data scientists cannot forget that after any inference task, they must take decisions regarding the final interpretation of the data.

**Acknowledgements** This chapter was co-written by Jordi Vitrià and Sergio Escalera.

## References

1. M.I. Jordan. Are you a Bayesian or a frequentist? [Video Lecture]. Published: Nov. 2, 2009, Recorded: September 2009. Retrieved from: http://videolectures.net/mlss09uk_jordan_bfway/
2. B. Efron, R.J. Tibshirani, *An introduction to the bootstrap* (CRC press, 1994)