- Misunderstandings of p-values on Wikipedia.
  https://en.wikipedia.org/wiki/Misunderstandings_of_p-values

- What does the 5 sigma mean?
  http://www.physics.org/article-questions.asp?id=103

9.8 Summary

In this tutorial, you discovered statistical hypothesis testing and how to interpret and carefully state the results from statistical tests. Specifically, you learned:

- Statistical hypothesis tests are important for quantifying answers to questions about samples of data.

- The interpretation of a statistical hypothesis test requires a correct understanding of p-values.

- Regardless of the significance level, the finding of hypothesis tests may still contain errors.

9.8.1 Next

In the next section, you will discover the three key distributions that you need to know well when working with statistical hypothesis tests.

Chapter 10

Statistical Distributions

A sample of data will form a distribution, and by far the most well-known distribution is the Gaussian distribution, often called the Normal distribution. The distribution provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space. This distribution describes the grouping or the density of the observations, called the probability density function. We can also calculate the likelihood of an observation having a value equal to or lesser than a given value. A summary of these relationships between observations is called a cumulative density function.

In this tutorial, you will discover the Gaussian and related distribution functions and how to calculate probability and cumulative density functions for each. After completing this tutorial, you will know:

- A gentle introduction to standard distributions to summarize the relationship of observations.
- How to calculate and plot probability and density functions for the Gaussian distribution.
- The Student's t and Chi-Squared distributions related to the Gaussian distribution.

Let's get started.

10.1 Tutorial Overview

This tutorial is divided into 4 parts; they are:

1. Distributions
2. Gaussian Distribution
3. Student's t-Distribution
4. Chi-Squared Distribution

10.2 Distributions

From a practical perspective, we can think of a distribution as a function that describes the relationship between observations in a sample space. For example, we may be interested in the age of humans, with individual ages representing observations in the domain, and ages 0 to 125 the extent of the sample space. The distribution is a mathematical function that describes the relationship of observations of different heights.

> A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.
>
> — Page 6, *Statistics in Plain English*, Third Edition, 2010.

Many data conform to well-known and well-understood mathematical functions, such as the Gaussian distribution. A function can fit the data with a modification of the parameters of the function, such as the mean and standard deviation in the case of the Gaussian. Once a distribution function is known, it can be used as a shorthand for describing and calculating related quantities, such as likelihoods of observations, and plotting the relationship between observations in the domain.

10.2.1 Density Functions

Distributions are often described in terms of their density or density functions. Density functions are functions that describe how the proportion of data or likelihood of the proportion of observations change over the range of the distribution. Two types of density functions are probability density functions and cumulative density functions.

- **Probability Density function**: calculates the probability of observing a given value.
- **Cumulative Density function**: calculates the probability of an observation equal or less than a value.

A probability density function, or PDF, can be used to calculate the likelihood of a given observation in a distribution. It can also be used to summarize the likelihood of observations across the distribution's sample space. Plots of the PDF show the familiar shape of a distribution, such as the bell-curve for the Gaussian distribution. Distributions are often defined in terms of their probability density functions with their associated parameters.

A cumulative density function, or CDF, is a different way of thinking about the likelihood of observed values. Rather than calculating the likelihood of a given observation as with the PDF, the CDF calculates the cumulative likelihood for the observation and all prior observations in the sample space. It allows you to quickly understand and comment on how much of the distribution lies before and after a given value. A CDF is often plotted as a curve from 0 to 1 for the distribution.

Both PDFs and CDFs are continuous functions. The equivalent of a PDF for a discrete distribution is called a probability mass function, or PMF. Next, let's look at the Gaussian distribution and two other distributions related to the Gaussian that you will encounter when using statistical methods. We will look at each in turn in terms of their parameters, probability, and cumulative density functions.

10.3 Gaussian Distribution

The Gaussian distribution, named for Carl Friedrich Gauss, is the focus of much of the field of statistics. Data from many fields of study surprisingly can be described using a Gaussian distribution, so much so that the distribution is often called the *normal* distribution because it is so common. A Gaussian distribution can be described using two parameters:

- `mean`: Denoted with the Greek lowercase letter mu (μ), is the expected value of the distribution.
- `variance`: Denoted with the Greek lowercase letter sigma (σ^2) raised to the second power (because the units of the variable are squared) , describes the spread of observation from the mean.

It is common to use a normalized calculation of the variance called the standard deviation

- `standard deviation`: Denoted with the Greek lowercase letter sigma (σ), describes the normalized spread of observations from the mean.

We can work with the Gaussian distribution via the `norm` SciPy module. The `norm.pdf()` function can be used to create a Gaussian probability density function with a given sample space, mean, and standard deviation. The example below creates a Gaussian PDF with a sample space from -5 to 5, a mean of 0, and a standard deviation of 1. A Gaussian with these values for the mean and standard deviation is called the Standard Gaussian.

```
# plot the gaussian pdf
from numpy import arange
from matplotlib import pyplot
from scipy.stats import norm
# define the distribution parameters
sample_space = arange(-5, 5, 0.001)
mean = 0.0
stdev = 1.0
# calculate the pdf
pdf = norm.pdf(sample_space, mean, stdev)
# plot
pyplot.plot(sample_space, pdf)
pyplot.show()
```

Listing 10.1: Example density line plot of Gaussian probability density function.

Running the example creates a line plot showing the sample space in the x-axis and the likelihood of each value of the y-axis. The line plot shows the familiar bell-shape for the Gaussian distribution. The top of the bell shows the most likely value from the distribution, called the expected value or the mean, which in this case is zero, as we specified in creating the distribution.

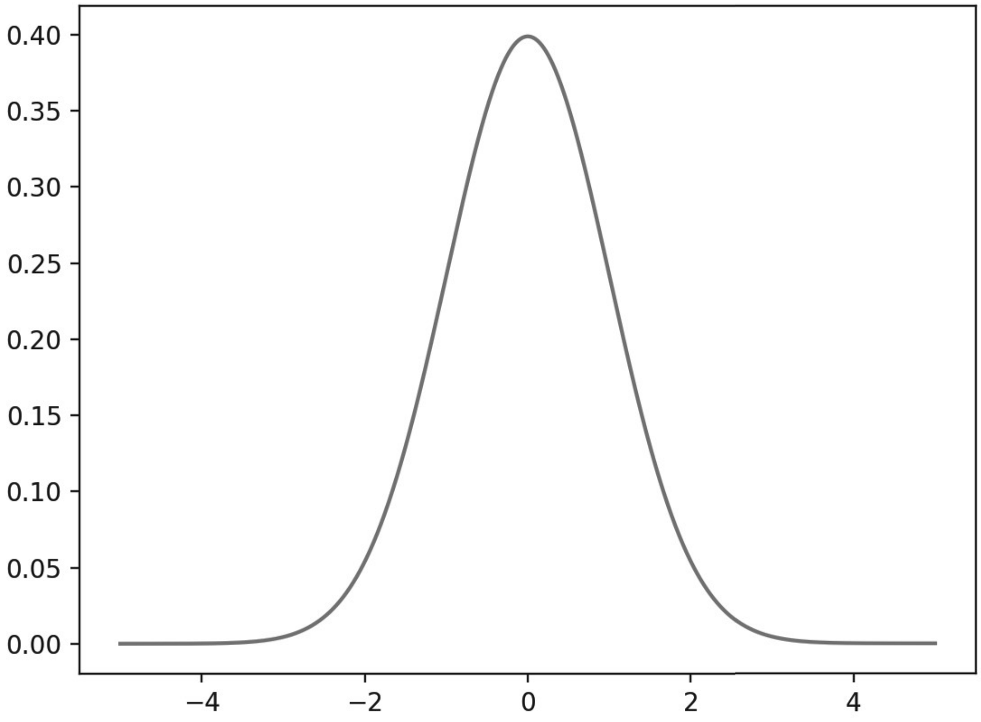

Figure 10.1: Density line plot of the Gaussian probability density function.

The `norm.cdf()` function can be used to create a Gaussian cumulative density function. The example below creates a Gaussian CDF for the same sample space.

```
# plot the gaussian cdf
from numpy import arange
from matplotlib import pyplot
from scipy.stats import norm
# define the distribution parameters
sample_space = arange(-5, 5, 0.001)
# calculate the cdf
cdf = norm.cdf(sample_space)
# plot
pyplot.plot(sample_space, cdf)
pyplot.show()
```

Listing 10.2: Example density line plot of Gaussian cumulative density function.

Running the example creates a plot showing an S-shape with the sample space on the x-axis and the cumulative probability of the y-axis. We can see that a value of 2 covers close to 100% of the observations, with only a very thin tail of the distribution beyond that point. We can also see that the mean value of zero shows 50% of the observations before and after that point.

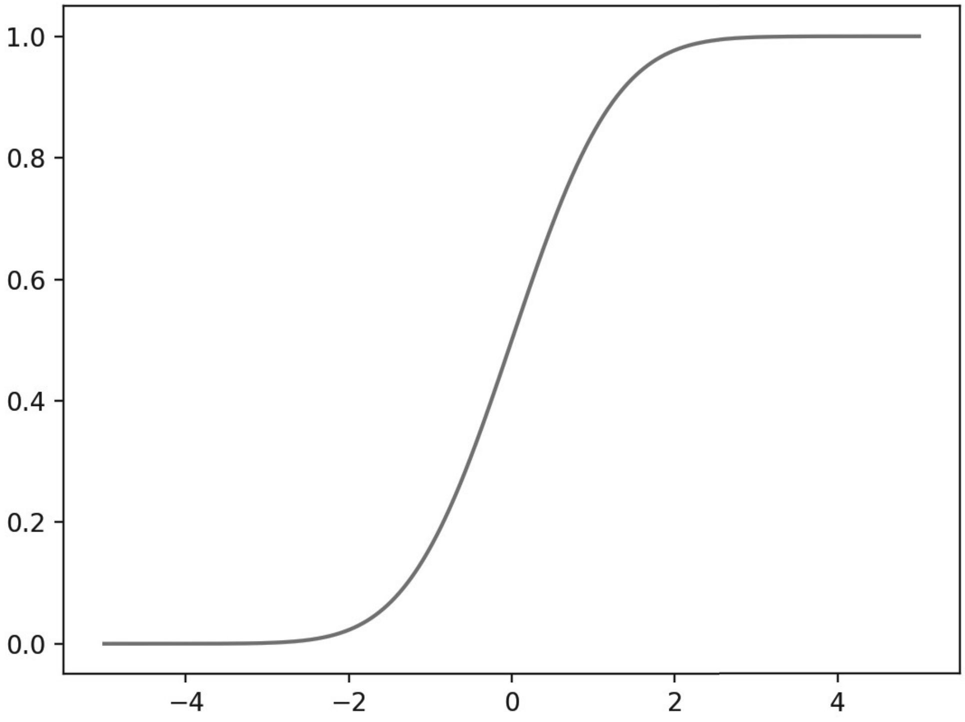

Figure 10.2: Density line plot of the Gaussian cumulative density function.

10.4 Student's t-Distribution

The Student's t-distribution, or just t-distribution for short, is named for the pseudonym *Student* by William Sealy Gosset. It is a distribution that arises when attempting to estimate the mean of a normal distribution with different sized samples. As such, it is a helpful shortcut when describing uncertainty or error related to estimating population statistics for data drawn from Gaussian distributions when the size of the sample must be taken into account.

Although you may not use the Student's t-distribution directly, you may estimate values from the distribution required as parameters in other statistical methods, such as statistical significance tests. The distribution can be described using a single parameter:

- `number of degrees of freedom`: denoted with the lowercase Greek letter nu (ν), denotes the number degrees of freedom.

Key to the use of the t-distribution is knowing the desired number of degrees of freedom. The number of degrees of freedom describes the number of pieces of information used to describe a population quantity. For example, the mean has `n` degrees of freedom as all `n` observations in the sample are used to calculate the estimate of the population mean. A statistical quantity that makes use of another statistical quantity in its calculation must subtract 1 from the degrees of freedom, such as the use of the mean in the calculation of the sample variance. Observations in a Student's t-distribution are calculated from observations in a normal distribution in order to describe the interval for the populations mean in the normal distribution. Observations are calculated as:

$$data = \frac{x - mean(x)}{\frac{S}{\sqrt{n}}} \tag{10.1}$$

Where x is the observations from the Gaussian distribution, $mean$ is the average observation of x, S is the standard deviation and n is the total number of observations. The resulting observations form the t-observation with $(n - 1)$ degrees of freedom. In practice, if you require a value from a t-distribution in the calculation of a statistic, then the number of degrees of freedom will likely be $n - 1$, where n is the size of your sample drawn from a Gaussian distribution.

> Which specific distribution you use for a given problem depends on the size of your sample.
>
> — Page 93, *Statistics in Plain English*, Third Edition, 2010.

SciPy provides tools for working with the t-distribution in the `stats.t` module. The `t.pdf()` function can be used to create a Student's t-distribution with the specified degrees of freedom. The example below creates a t-distribution using the sample space from -5 to 5 and (10,000 - 1) degrees of freedom.

```
# plot the t-distribution pdf
from numpy import arange
from matplotlib import pyplot
from scipy.stats import t
# define the distribution parameters
sample_space = arange(-5, 5, 0.001)
dof = len(sample_space) - 1
# calculate the pdf
pdf = t.pdf(sample_space, dof)
# plot
pyplot.plot(sample_space, pdf)
pyplot.show()
```

Listing 10.3: Example density line plot of Student's t probability density function.

Running the example creates and plots the t-distribution PDF. We can see the familiar bell-shape to the distribution much like the normal. A key difference is the fatter tails in the distribution (hard to see by eye), highlighting the increased likelihood of observations in the tails compared to that of the Gaussian.