ST101 Unit 5: Inference

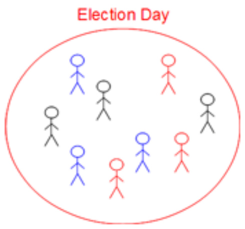
Contents:

Confidence Intervals

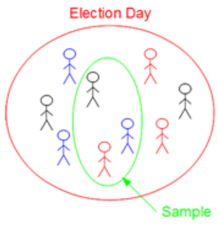
In this section we will talk about confidence intervals, and later in this unit we will look at testing hypotheses. These are fundamental concepts that every applied statistician uses almost every day.

To introduce the concept of the confidence interval, let's look at an example.

Imagine that there is an election. As a statistician, you will often want to try and understand what the outcome of the election will be before the actual vote happens:



Of course, you could just go out and ask every voter, but the effort involved would be almost the same as running the entire election in the first place! Although this may be possible for a small group of people, say 5 or 10 perhaps, but for an electorate numbering in the tens, or even hundreds of millions it is just not feasible.



What statisticians actually do is to choose a random sample and hope that the responses from this sample are representative of the whole group.

The sample will be a (hopefully) randomly drawn sample of people from the pool of voters who we then assume to be representative of the pool at large.

Using the sample, the statisticians will come up with an estimate of how they expect people to vote. These estimates are often reported along the lines of:

> "*60% will vote for party A, and 40 % for party B*"

In practice however, the statisticians actually report back their estimate together with a margin of error, for example:

Party A:	60% ±3%
Party B:	40% ±3%

This margin of error means that what is being returned isn't just a single number like 60%, but rather a range or interval. We call this the **confidence interval**. In the case of Party A in the example above, the lower bound of the confidence interval is 57%, and the upper bound is 63%.

What this means is that, given our current sample, we are fairly confident that, come election day, Party A will achieve an outcome within the range 57-63%.

Confidence intervals are a fundamental concept in statistics. They can be applied to coin flips and many of the other examples we saw earlier.

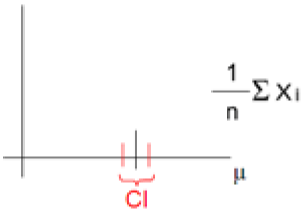
A candidate standing in an election may have a true chance, p, that any given voter will vote for him. Now, if $p > 0.5$ in a two-candidate run-off election, then he will win the election in most cases (although not always).

As statisticians, however, we cannot assess the true chance. What we do is we form a sample group:

$$X_1 \ldots X_n$$

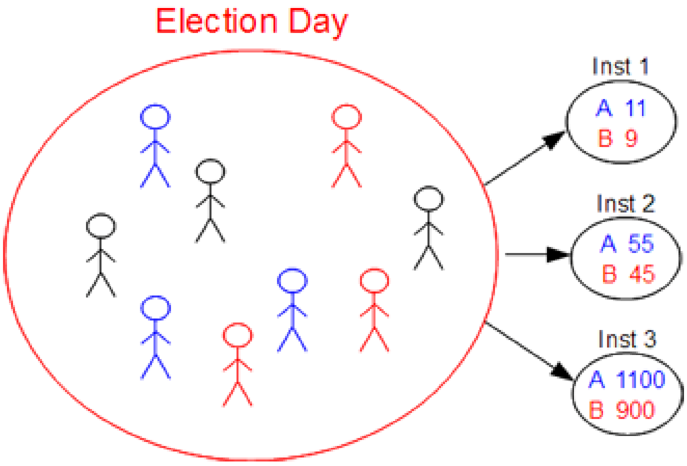
In coin-flipping, we flip n coins; in an election we ask n people chosen at random. The results from this sample group will give us an estimated mean, μ, and estimated variance, σ^2 (which then gives us the standard deviation).

What is new is the confidence interval. The confidence interval, CI, is not the same as the variance. What we are saying when we quote the confidence interval is that, based on the outcome with the sample group, we believe that the parameter μ (usually based on the MLE) falls within the range given by the CI:



Very often, the CI is defined as the range where there is a 95% chance that the outcome will occur. So how do we compute the confidence interval?

Let's stick with our election example, and for simplicity we will say that there are only two parties. Let's also say that there are three institutions sampling the voters to try to predict the election result. The first institution samples 20 voters, the second institution samples 100 voters and the third institution samples 1000 voters:



In each case, using the maximum likelihood estimator, the probability that a voter will vote for Party A, P(A), is 0.55. Clearly, we would have more confidence in the prediction from institution 3 because they used a much larger sample size.

Consider an extreme case where a company only sampled one voter. If that voter said that they intended to vote for Party A, then the company will be forced to conclude that P(A) = 1. Would we trust this prediction? Of course not! In general, the more data that is sampled (assuming that the sampling is fair and independent), the more trust we will have in the result. More trust means a smaller confidence interval.

So, if we increase the sample size, N, the confidence interval will get smaller. By contrast, the standard deviation will be unchanged if we increase the sample size. The standard deviation distribution is not dependent on the sample.

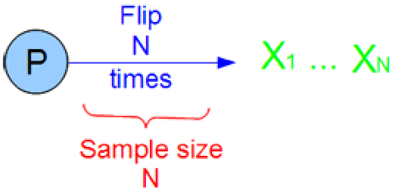
NOTE: Standard Deviation in this case means the standard deviation of the population from which the elements are sampled.

This is calculated as: $\sum_{i=0}^{n} \frac{(x_i - \mu)^2}{n}$

In fact, it turns out that $CI \propto \frac{\sigma}{\sqrt{N}}$

Let's go back probability, and to our simple coin-flip example.

We'll assume that the probability of the coin coming up heads in P, and that we flip the coin N times. This gives us a sample set of size N:



In the past, we have calculated the empirical mean, μ, and the variance, σ^2. Now we are going to calculate the confidence interval, CI.

We know that if the probability of heads P(H) = 0.5 then the expected mean will be:

$$\mu = 0.5$$

We also know that the variance is defined as the quadratic difference of the sum of the differences of the outcomes from the mean:

$$\sigma^2 = \frac{1}{N}\sum(X_i - \mu)^2$$

In this case, there are two possible outcomes, heads, 1, or tails, 0. Both outcomes have a probability of 0.5.

If the outcome is 1, the squared difference from the mean is:

$$(1-0.5)^2 = 0.25$$

Similarly, if the outcome is 0, the squared difference from the mean is:

$$(0-0.5)^2 = 0.25$$

So the variance is:

$$\frac{(1-0.5)^2+(0-0.5)^2}{2} = 0.25$$

Now, let's say we have three sample groups of size 1, 2, and 10. We know that the mean, μ, is 0.5, and that the variance of each individual coin-flip, σ^2, is 0.25. We saw that when we add Gaussian variables, the means and variances add up, so we can calculate the mean and variance of the sums of all the outcomes as follows:

	MEAN(ΣX_i)	VAR(ΣX_i)
N = 1	0.5	0.25
N = 2	1	0.5
N = 10	5	2.5

We can also calculate the variance of the means, $VAR\left(\frac{1}{N}\sum X_i\right)$.

Since variance is a quadratic expression,

$$VAR(aX) = a^2 VAR(X)$$

and so we have:

	MEAN(ΣX_i)	VAR(ΣX_i)	$VAR\left(\frac{1}{N}\sum X_i\right)$
N = 1	0.5	0.25	$\frac{1}{N}VAR\left(\sum X_i\right) = \frac{1}{1}\times 0.25 = 0.25$
N = 2	1	0.5	$\frac{1}{N}VAR\left(\sum X_i\right) = \frac{1}{2}\times 0.5 = 0.125$
N = 10	5	2.5	$\frac{1}{N}VAR\left(\sum X_i\right) = \frac{1}{10}\times 2.5 = 0.025$

This tells us something quite profound. The variance of the sum, VAR(ΣX_i), increases as the sample size, N, increases. However, the variance of the mean, or *spread of the mean*, actually decreases as the sample size increases.

As we already know, the standard deviation of the means will be just the square root of the variances of the means, thus:

	MEAN(ΣX_i)	VAR(ΣX_i)	$VAR\left(\frac{1}{N}\sum X_i\right)$	$SD\left(\frac{1}{N}\sum X_i\right)$
N = 1	0.5	0.25	0.25	0.5
N = 2	1	0.5	0.125	0.354
N = 10	5	2.5	0.025	0.05

Now we are able to calculate the confidence interval. It turns out that if we multiply the value we just calculated for the standard deviation of the mean by 1.96 we get the confidence interval for that sample size:

	MEAN(ΣX_i)	VAR(ΣX_i)	$VAR\left(\frac{1}{N}\sum X_i\right)$	$SD\left(\frac{1}{N}\sum X_i\right)$	C.I.
N = 1	0.5	0.25	0.25	0.5	0.98
N = 2	1	0.5	0.125	0.354	0.686
N = 10	5	2.5	0.025	0.158	0.3136

Now a note of caution. This trick of multiplying by 1.96 to calculate the confidence interval isn't mathematically correct for very small sample sizes. It normally assumes that we have at least 30 samples.

Confidence at 100 Quiz

Calculate the values for a sample size N = 100.

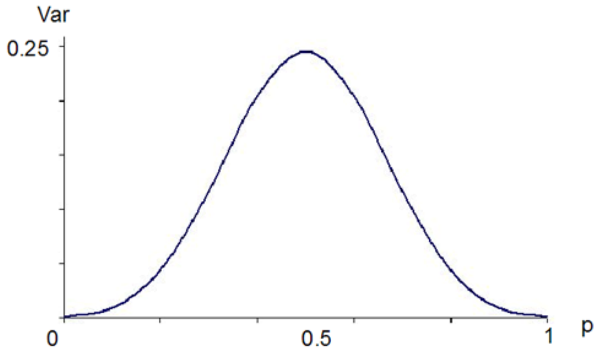
Variance Formulae

Now, we have studied the special case of a fair coin where p = 0.5, let's look at the more general case for an arbitrary value of p.

It turns out that the real challenge isn't calculating the confidence interval, but actually calculating the variance, σ^2. Let's consider the extreme cases where p = 0 or p = 1.

When p = 0, the coin always comes up tails, and the variance is therefore 0. Similarly, when p = 1, the coin always comes up heads, and the variance is also 0.

Now, we already know that when p = 0.5 the variance $\sigma^2 = 0.25$ and if we were to plot variance against p, we would expect to see something like this:



In fact, the formulae for calculating the mean and variance for a coin where P(H) = p are:

$$\mu = p$$
$$\sigma^2 = p(1 - p)$$

Now, given that p is the mean of X, we can derive the formula for the variance as follows.

There are two possible outcomes for the coin-flip. If the coin comes up heads, the variance is the product of p times the quadratic difference of 1 (for heads) minus the mean, p:

$$p(1 - p)^2$$

If the coin comes up tails, we get a similar result with 0 for tails:

$$(1 - p).p^2$$

Therefore:

$$\text{Var}(X) = p.(1 - p)^2 + (1 - p).p^2$$

Multiplying this out gives:

$$\begin{aligned}\text{Var}(X) &= p(1 - 2p + p^2) + p^2 - p^3 \\ &= p - 2p^2 + p^3 + p^2 - p^3 \\ &= p - p^2 \\ &= p(1 - p)\end{aligned}$$

CI Quiz

You have a loaded coin where p = 0.1

Calculate the variance Var(p) and the confidence intervals for sample sizes:

- N = 1
- N = 10
- N = 100

Use the following formula to calculate CI: $1.96\sqrt{\frac{(1-p)}{N}}$

CI Quiz 2

Suppose you flip a coin and observe the following sample set:

0, 1, 1, 1

Calculate the mean, the variance, and the confidence interval. You should calculate the actual variance from the data sequence, not the variance given by the formula.

Note: You should use the formula:

$$1.96\sqrt{\frac{\sigma^2}{N}}$$

to calculate the confidence interval, even though the sample size, N, is only 4.

In practice, you would only use this formula for samples where N ≥30

CI Quiz 3

Now calculate the mean, the variance, and the confidence interval for the sample set:

0, 0, 0, 1, 1, 1, 1, 1, 1, 1

CI Quiz 4

What happens to the mean, the variance, and the confidence interval if we flip the coin 1000 times and get the following result:

- 400 x tails
- 600 x heads

Normal Quantiles

In the last section we introduced the “magic number”, 1.96, as the multiplier for calculating the confidence interval. We defined the confidence interval to be:

$$\text{Mean} \ \pm \ 1.96\frac{\sigma}{\sqrt{N}}$$

where N is the number of samples. But where did the “magic” number 1.96 come from?

You will recall that when we looked at the central limit theorem we found that, when we have a large sample size, N, the mean outcome for N independently drawn values, X_i

$$\text{Mean} = \frac{1}{N}\sum X_i$$

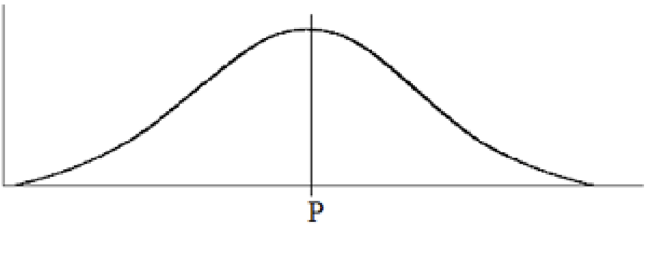
becomes more and more normal. It runs out that the “magic” number, 1.96, is directly related to this finding from the central limit theorem.

Take our coin-flip example. The coin has a true probability, P, but we cannot measure this. What we can do is to take our sample of N flips and calculate the empirical mean using the formula:

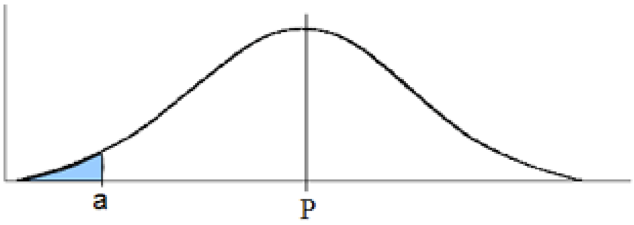
$$\mu = \frac{1}{N}\sum X_i$$

Now, we have already seen that it is quite likely that $p \neq \mu$ since the best that we can hope for from a finite number of coin-flips is just an estimate of the true probability, P.

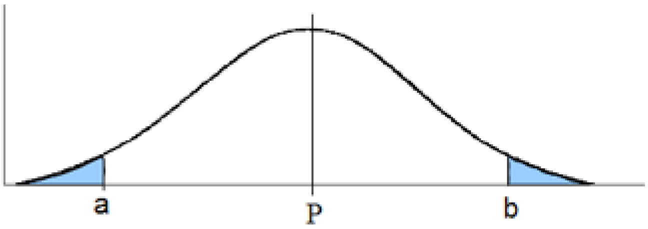
We know that for large sample sizes ($N \geq 30$) the distribution of μ that you might observe is Gaussian:



Now, the probability of observing any particular value of μ is the height of the curve at that point. so, for any value a, the chance of observing a value of μ smaller than a, P(μ < a), is given by the area under the curve as shown:



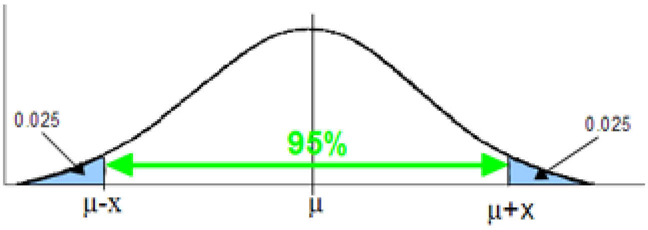
Also, by symmetry, the chance of observing a value of μ > b, where b is the same distance from p as a, is given by the area under the curve to the right of b:



If we compute the value of x such that when

$a = \mu - x$ and $b = \mu + x$

we have exactly 2.5% of the area under the curve enclosed on either side, then we can be 95% certain that μ will be between these limits:



This **95% confidence interval** occurs when x = 1.96. Making x smaller reduces the size of the confidence interval, while increasing x increases the size of the confidence interval:

90% confidence interval	x = 1.64
99% confidence interval	x = 2.58

The values for x are called **quantiles**.

T-Tables

If we have less than 30 samples, we are likely to experience problems if we try to use quantiles to calculate our confidence intervals. In this situation we can use a tool called a T-table. These are listed in most statistical textbooks.

Here is a selection of values from a T-table:

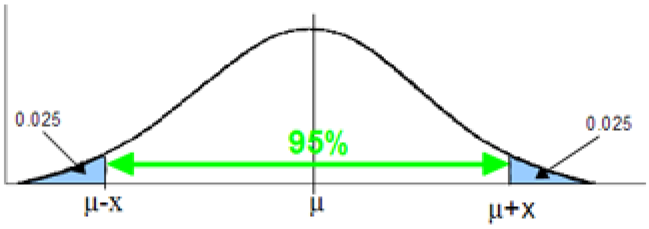
	Significance Level = α					
Degrees Of Freedom	**0.005(1 tail)**	**0.01(1 tail)**	**0.025(1 tail)**	**0.05(1 tail)**	**0.10(1 tail)**	**0.25(1 tail)**
	0.01(2 tails)	**0.02(2 tails)**	**0.05(2 tails)**	**0.1(2 tails)**	**0.20(2 tails)**	**0.50(2 tails)**
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	0.816
3	5.841	4.541	3.182	2.353	1.638	0.765
4	4.604	3.747	2.776	2.132	1.533	0.741
5	4.032	3.365	2.571	2.015	1.476	0.727
6	3.707	3.143	2.447	1.943	1.440	0.718
7	3.500	2.998	2.365	1.895	1.415	0.711
8	3.355	2.896	2.306	1.860	1.397	0.706
9	3.250	2.821	2.262	1.833	1.383	0.703
10	3.169	2.764	2.228	1.812	1.372	0.700
11	3.106	2.718	2.201	1.796	1.363	0.697
12	3.054	2.681	2.179	1.782	1.356	0.696
13	3.012	2.650	2.160	1.771	1.350	0.694
14	2.977	2.625	2.145	1.761	1.345	0.692
15	2.947	2.602	2.131	1.753	1.341	0.691
16	2.921	2.584	2.120	1.746	1.337	0.690
17	2.898	2.567	2.110	1.740	1.333	0.689
18	2.878	2.552	2.101	1.734	1.330	0.688
19	2.861	2.540	2.093	1.729	1.328	0.688
20	2.845	2.528	2.086	1.725	1.325	0.687

This may need some explanation!

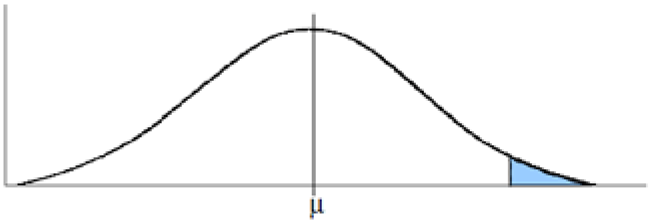
For our purposes, if we have N samples, the **Degrees of Freedom** is (N – 1). So, if our sample set contained 17 samples we would look to the row where Degrees of Freedom is 16.

The numbers in the header row are the (1 – CI), so if you want the 95% confidence interval you look to the column headed (1 – 0.95) = 0.05.

You will have noticed that the header row is divided in two with values for “1-tail” and also for “2-tails”. So far, we have been considering the “2-tail” case where we cut-off to both left and right of our confidence interval:



There are occasions (we will meet these later), where we only want to cut off one side:



These often occur in the context of testing hypotheses, but for the time being, we can limit ourselves to the “2-tail” row.

The t family of distributions arises in a case where the underlying population is approximately normally distributed (e.g. heights) and both the mean and variance must be estimated. The additional uncertainty in the variance results in thicker tails than the normal distribution. For larger numbers of degrees of freedom the t-distribution becomes increasingly similar to the normal distribution.

You can find more about the use of the t-distribution vs the normal at Khan Academy and Wikipedia.

So, as an example, if we have 8 samples in our sample set, and we want the 90% confidence interval we would look in the row where Degrees of Freedom = 7, and the column headed “0.1 (2-tails)” and locate the value 1.895.

Reading Tables Quiz

Suppose we want to achieve a 95% confidence interval, and we want a factor that is:

≥ 2.415

What is the minimum number of data points we need to collect to achieve this?

Reading Tables Quiz 2

Let’s say that we flip a coin 5 times and get the sequence:

1, 1, 0, 0, 0

Calculate the following values:

- σ^2
- $\sqrt{\frac{\sigma^2}{N}}$
- CI

The confidence interval should be given in the form :

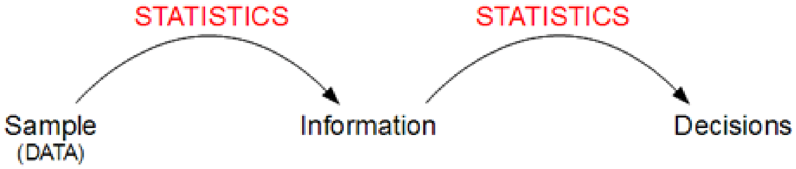
$$\text{MEAN} \pm x\sqrt{\frac{\sigma^2}{N}}$$

where x is the value from the T-table.

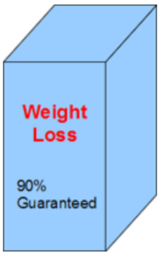
Note that in practice the t-distribution is not used to estimate binomial probabilities. Since the underlying distribution is not normal, on small samples exact confidence intervals are computed as described in the next section.

Hypothesis Test

Hypothesis testing is really all about decision making. So far on this course, we have talked about data a lot, and we have used statistics to extract information from that data. Now, we are going to use more statistics to make decisions based on that information:



These decisions will be binary, that is the outcome will be either ‘yes’ or ‘no’.



Let’s start with an example. Imagine that you have been given a carton of weight-loss pills. On the box it says that if you take these pills a substantial weight loss is guaranteed in 90% of cases.

It is your job to try to establish whether or not the pill manufacturer’s claim is accurate. You contact people who took the pills and asked the question “Did you lose weight?”.

You asked 15 people, and all of them answered. Eleven of these people answered ‘Yes’, while 4 people answered ‘No’. Now this is only a 73.3% success rate, which is far lower than the manufacturer’s claim on the carton. However, as we now know, this could just be a sample error. We will use hypothesis testing to see whether we should accept the manufacturer’s claim on the basis of our data, or whether we should contact the manufacturer to complain that their claim is incorrect.

In this case we start with the hypothesis that the manufacturer’s claim is correct. This starting hypothesis is often called the **Null-hypothesis**, H_0. We also have a counter hypothesis - one that invalidates the Null-hypothesis. This is called the **Counter-hypothesis**, H_1.

In this case, we can write:

$$H_0: p = 0.9$$

$$H_1: p < 0.9$$

Now we have these two hypotheses, and we are going to test them. We accept the Null-hypothesis until it is proved to be wrong, and the Counter-hypothesis is correct (i.e. we have sufficient evidence to show that there is a very high likelihood that the Null-hypothesis is wrong).

Critical Region

In the case of our weight-loss pills, the result from our sample group was:

YES: 11

NO: 4

and our hypotheses are:

H_0: $p = 0.9$

H_1: $p < 0.9$

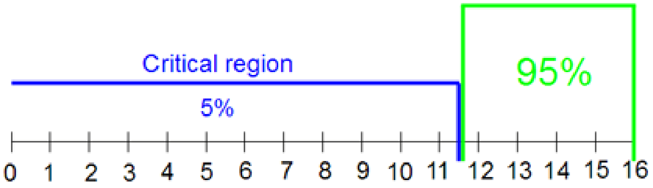
Now this sample set is obviously binomial ('binomial' just means binary results with just two possible outcomes, like coin-flips), and with this sample set there are exactly 16 possible outcomes, from nobody saying they lost weight, to all 15 members of the sample group saying that they had lost weight. We can calculate the probabilities for the binomial distribution using:

$$\frac{N!}{k!(N-k)!}p^k(1-p)^{(N-k)}$$

we get the result:

0	0	0	0	0	0	0	0	0.0003	0.002	0.01	0.04	0.13	0.27	0.34	0.21
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

What we want to do is to identify the group of outcomes on the left of this distribution that total no more than 5% of the total probability (for a 95% confidence level). Everything in this group is called the **critical region**. Everything to the right of this group is the **acceptance region**:



Outcomes in the critical region invalidate the Null-hypothesis. Outcomes in the acceptance region validate it.

In our case, summing all the values from 0 to 10 gives a result less than 0.05 (5%), but adding the value for 11 causes the result to exceed this threshold. The critical region in this case is therefore 0 – 10.

CRITICAL REGION											ACCEPTANCE REGION				
0	0	0	0	0	0	0	0	0.0003	0.002	0.01	0.04	0.13	0.27	0.34	0.21
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Loaded Coin Example

Let’s work through another example. Last weekend, Sebastian bought a loaded coin from the Magic Shop. He was told that for his new coin

P(HEADS) = 0.3

Sebastian suspects that the shop sold him a fair coin (or something close to a fair coin), so this is a 1-sided hypothesis.

Obviously, in this case the Null-hypothesis will be:

H_0: p = 0.3

and the Counter-hypothesis will be:

H_1: p > 0.3

Sebastian flipped the coin eleven times and got the sample set:

T, H, H, T, T, T, H, T, T, H, H

The observed value for p is therefore:

$$p = \frac{5}{1} = 0.45$$

Should Sebastian return the coin, given that the observed probability for this sample set is much higher than the advertised probability of 0.3?

Once again, we can calculate the binomial probability distribution using:

$$\frac{N!}{k!(N-k)!} p^k (1-p)^{(N-k)}$$

and we get:

0.02	0.09	0.12	0.25	0.22	0.13	0.06	0.02	0	0	0	0
0	1	2	3	4	5	6	7	8	9	10	11

Now, we can see that the most-likely outcome for 11 coin flips is 3 Heads, and the critical region will be somewhere to the right of 3 in this table.

In fact, for a 95% confidence level, the critical region is the columns 7 to 11 in the table, since including column 6 would push the total probability over 0.05.

So the observed five heads out of eleven flips falls well within the safe region. Sebastian would only need to return the coin if he had seen seven or more heads.

Fair Coin Example

So now we go to a bank to get what we hope is a fair coin. Our Null-hypothesis is thus:

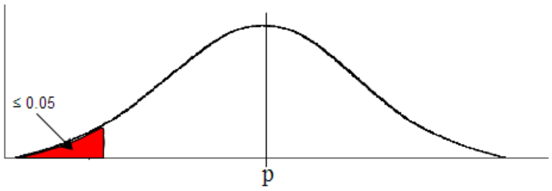
H_0: p = 0.5

and the Counter-hypothesis will be:

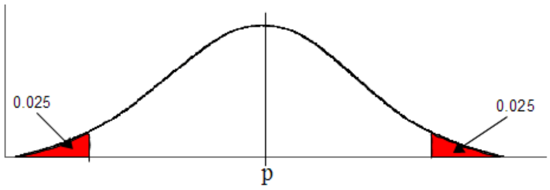
H_1: p $\neq$ 0.5

Note that this time we have a 2-sided hypothesis. The probability may be smaller than 0.5 or higher than 0.5.

The way to look at this conceptually is that all we have done so far is to assume that H0 is correct, computed some sort of distribution, and then cut out a critical region such that the area under the curve did not exceed 0.05, or 5% (assuming we wanted 95% confidence):



In the 2-sided test, we cut out a smaller region on the left of the curve, but also one on the right, such that the area under the curve on each side does not exceed 0.025 or 2.5% for 95% confidence:



The total area under the curve (both left and right) still does not exceed 5%.

This is called a **2-tailed test**, and you’ll have noticed that it looks an awful lot like the confidence interval!

So now we flip the coin. In 14 flips we got the following results:

T, T, T, H, H, T, H, T, T, T, T, T, T, T

So let’s do the analysis.

In this table, we've listed the probabilities calculated for the binomial distribution:

0	0	0.005	0.022	0.06	0.12	0.18	0.21	0.18	0.12	0.06	0.022	0.005	0	0
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

The critical region in this case is going to be those columns on the left where the sum of the probabilities is ≤ 0.025, and the columns on the right where the sum of the probabilities is ≤ 0.025. i.e. columns 0 – 2, and columns 12 – 14 as shown:

Critical region			Acceptance Region									Critical region		
0	0	0.005	0.022	0.06	0.12	0.18	0.21	0.18	0.12	0.06	0.022	0.005	0	0
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

So it appears that our 3 heads falls within the acceptance region for our coin.

Cancer Treatment

A treatment for cancer is advertised. The manufacturer claims that it works in 80% of all cases. You suspect that the drug may not work as well as advertised.

Suppose that ten people are treated with the drug. The most likely outcome, if the manufacturer's claims are true is that eight of them with be cured.

Cancer Treatment Quiz

Using the 95% confidence level, what is the largest number of healthy people in the critical region that would cause you to reject the manufacturer's claims for this drug?

Cancer Treatment Quiz 2

In the experiment, 5 out of the 10 subjects in the sample set were found to be healthy, and five still had cancer.

Do we reject the manufacturer's claim that the treatment will work in 80% of all cases?

Summary

Congratulations! You now understand the basics of hypothesis testing. You should understand what is meant by the terms **critical region** and **null-hypothesis**, and you have seen how to apply them in both a 1-sided and 2-sided test.

Essentially, if the actual outcome observed from our sample set falls into the critical region then we will become suspicious and reject the null hypothesis. Whereas, if the observed outcome falls into the 95% acceptance region then we would accept the hypothesis as valid. That is the essence of hypothesis testing.

Hypothesis Test 2

In this section, we will combine what we learned in the last section with what we learned about confidence intervals.

Height Test Example

According to Wikipedia, in the 2007-2008 season, the average NBA basketball player was 6’ 6.98” (200 cm) tall. Let’s question this. Is this acceptable?

We visit an NBA training game and measure a group of players. We observe the following heights:

199cm, 200cm, 201cm, 202cm, 203cm, 204cm, 205cm, 206cm

On the basis of these measurements, and with a 95% confidence level, should we reject the claim in Wikipedia?

We have already seen how to calculate confidence intervals using the, now familiar, formula:

$$\frac{1}{N}\sum X_i \quad \pm a\sqrt{\frac{\sigma^2}{N}}$$

Where a is the factor from the T-table. In this case, a = 2.365, and if we plug our observed data into the formula we get:

202.5 ± 1.916

This means that the 95% confidence interval stops at 200.58cm. The figure of 200cm from Wikipedia therefore falls into the critical region, and we should reject the 200cm hypothesis based on our sample of eight people.

Now this would not be correct in practice. In fact, 6’ 6.98” = 200.61cm, which falls within the confidence interval. The figure was rounded to 200cm for the purposes of this example. Also, we might find a very different sample if we were to visit a different town. The sampling wasn’t completely independent.

However, the principle illustrated here is valid. Hypothesis testing can be really simple if we use our confidence interval and check whether the observed outcome lies within or outside that confidence interval.

We now know how to compute the confidence interval for any sample. If our null-hypothesis H0 falls inside the confidence interval then we accept that hypothesis. If, on the other hand, the null-hypothesis falls outside the confidence interval, then we reject the null-hypothesis and we accept the alternate-hypothesis.

In summary, given a sample of data, X1 … XN, we the calculate the mean and variance using:

$$\mu = \frac{1}{N}\sum X_i$$

$$\sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2$$

We then get the T-value, at some desired error probability, p, from the tables:

T(N-1, p) = a

Remembering to select the correct T-value according to whether we are considering a 1-sided or a 2-sided test. Then the ± term in the confidence interval is simply:

$$a\sqrt{\frac{\sigma^2}{N}}$$

The lower bound of the confidence interval will therefore be $\mu - a\sqrt{\frac{\sigma^2}{N}}$ and the upper bound will be $\mu + a\sqrt{\frac{\sigma^2}{N}}$

NOTE: You should be aware that in other statistics courses, and for smaller samples, the variance is computed using N-1 as the divisor instead of n. So:

$$\sigma^2 = \frac{1}{N-1}\sum (X_i - \mu)^2$$

Club Age

A dance club operator advertises that the average age of its clients is 26. You visit the club and encounter 30 people with the following age distribution:

Number of people	Age
4	21
6	24
7	26
11	29
2	40

Do you trust the club operator’s claim, based on this data sample?

Calculating the mean and the variance is straightforward, and we get:

$$\mu = 28.97$$

$$\sigma^2 = 19.57$$

Club Age Quiz

Noticing that N = 30, what value of a will we use for a 2-tailed, 95% confidence level?

Club Age Quiz 2

What is the ± term in the confidence interval?

So there is no real reason to doubt the club operator’s claim based upon the sample that we have drawn.

Once again, in reality we would need to be aware that on any given night there might be a reason why particularly young people come to the club, or that on another night older people might come, perhaps due to the style of music being played. As a statistician, we would be wary of judgements made based on a sample taken on a single night. To ensure that the sample is truly independent you would need to attend on random nights, and pick a random person each night.

It’s a tough life being a statistician!

Programming Tests and Intervals (Optional)

This is the optional programming section.

What we want here is really simple.

We want code that takes a sample and a hypothesis and returns either ‘Yes’ or ‘No’ depending on whether the result falls within the confidence interval.

For simplicity, we will assume 95% confidence and a 2-sided test.

Confidence Intervals Quiz

Write a function conf() that computes the ± term in the confidence interval. Use the functions mean() and var() that we wrote earlier as needed.

Hypothesis Test Quiz

Write a function test(), that takes as input a sample, l, and a hypothesis, h, and returns True if we believe the hypothesis in a 2-sided test, and otherwise returns False.

Answers

Confidence at 100 Quiz

	MEAN(ΣX_i)	VAR(ΣX_i)	$VAR\left(\frac{1}{N}\sum X_i\right)$	$SD\left(\frac{1}{N}\sum X_i\right)$	C.I.
N = 100	50	25	0.0025	0.05	0.098

CI Quiz

Var(p) = 0.09

- N = 1 CI = 0.588
- N = 10 CI = 0.186
- N = 100 CI = 0.0588

CI Quiz 2

$\mu = 0.75$

$$\sigma^2 = \frac{(0-0.75)^2 + (1-0.75)^2 + (1-0.75)^2 + (1-0.75)^2}{4} = 0.185$$

CI = 0.424

CI Quiz 3

$\mu = 0.7$

$$\sigma^2 = \frac{3\times(0-0.7)^2 + 7\times(1-0.7)^2}{10} = 0.21$$

CI = 0.284

CI Quiz 4

$\mu = 0.6$

$$\sigma^2 = \frac{400\times(0-0.6)^2 + 600\times(1-0.6)^2}{1000} = 0.24$$

CI = 0.0304

Reading Tables Quiz

15

Reading Tables Quiz 2

- σ2 = 0.24
- $\sqrt{\frac{\sigma^2}{N}} = 0.219$
- CI = 0.4 ± 0.467

Cancer Treatment Quiz

This is clearly a 1-sided test (we're not concerned if the test works better than advertised).

H_0 p = 0.8
H_1 p < 0.8

The binomial distribution looks like this, with the critical region shown:

0	0	0	0	0.005	0.026	0.088	0.2	0.3	0.27	0.1
0	1	2	3	4	5	6	7	8	9	10

So the largest number of healthy people in the critical region 1s 5.

Cancer Treatment Quiz 2

Yes.

Club Age Quiz

1.96

Note: While this is not required for this course, you may want to know that 1.96 is based on the assumption that the sampling distribution is close enough to normal.

The actual value is from a t-distribution with 29 degrees of freedom and is closer to 2.045.

Club Age Quiz 2

$$1.96\sqrt{\frac{19.57}{30}} = 1.58$$

Confidence Intervals Quiz

```
from math import sqrt

def mean(l):
   return float(sum(l))/len(l)

def var(l):
   m = mean(l)
   return sum([(x-m)**2 for x in l])/len(l)

def factor(l):
   return 1.96


def conf(l):
   return factor(l) * sqrt(var(l)/len(l))
```

Hypothesis Test Quiz

```
from math import sqrt

def mean(l):
   return float(sum(l))/len(l)

def var(l):
   m = mean(l)
   return sum([(x-m)**2 for x in l])/len(l)

def factor(l):
   return 1.96


def conf(l):
   return factor(l) * sqrt(var(l) / len(l))


def test(l, h):
   m = mean(l)
   c = conf(l)
   return abs(h - m) < c
```