Chapter 6. Statistical Inference

When one statistically infers, one is drawing conclusions from the sample about a population. *Statistical inference* also states how confident we are of our conclusions, expressed in probabilities. This chapter concerns only the setting: Inference about the mean of a Normal population, whose standard deviation is known (Note: this is actually unrealistic). This chapter involves two types of formal inference: *Confidence intervals* and *Tests of Significance*. These both state probabilities that would occur after many repetitions of the inference method. Another important point to mention is that statistical inference is most reliable when a properly randomized sample is used when gaining the data.

6.1: Confidence intervals

In repeated sampling the sample mean $\bar{x}$ is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$. The 68-95-99.7 rule states that $\bar{x}$ is within 2σ of μ 95% of the time, so about 95% of all samples will contain the true μ in the interval from $\bar{x}$ -2$\sigma_{(\bar{x})}$ and $\bar{x}$ +2$\sigma_{(\bar{x})}$. This $\sigma_{(\bar{x})}$ is equal to $\frac{\sigma}{\sqrt{n}}$.

For example, say we know that σ=100 (this is not realistic as we would not know the standard deviation upfront), n=500 and we have a $\bar{x}$ of a certain sample that is 470, then 2$\sigma_{(\bar{x})}$ =$\frac{100}{\sqrt{500}}$ = 4.5. Therefore we are 95% confident that the unknown μ is between 470 – (2 x 4.5) = 461 and 470 + (2 x 4.5) = 479. There is a 5% probability that the μ could be above 479 and below 461. The meaning of these statements is that 95% of the time this method gives a correct answer. The interval of numbers between $\bar{x}$ ± 2$\sigma_{(\bar{x})}$ is called the *confidence interval* for μ. Another way to see this is: Estimate ± Margin of Error. The *margin of error* gives us an idea of how accurate our estimation is.

A confidence interval can be chosen, however the 95% interval is used most. C stands for the confidence interval in decimal form; therefore in this case, C is 0.95.
A more formal manner to explain the margin of error for a level C confidence interval is through the formula m = $z^* \cdot$. On a standard Normal curve, the z* is the value of the critical points –z* and z*, with area C within these two points. The level C confidence interval for μ is therefore $\bar{x}$ ± m. When n is large and the distribution is normal then the interval is exact. To find the value of z* for different confidence levels see Table D.

Changes in the sample sizes change the margin of error, for example by decreasing the sample size to ¼ of its original value; the margin of error is doubled. High confidence and small margins of error are desirable, so what can we do if a margin of error is too big? If you look at the formula for m, there are 3 things you can change:

- You can use a lower confidence level (smaller C and therefore a smaller z*).
- You can increase the sample size (larger n)
- You can reduce the standard deviation, by carefully controlling the measurement process or by focussing only on a portion of a large population.

Also by looking at the formula for m, you can see that due to the squar e root in the formula, one would have to multiply the sample size by 4 to make the margin of error half of the original value.

When planning a data collection, a user of statistics can specify n when wanting a certain margin of error by using the formula $n = {}^{2}$. When the result of this formula is not a whole number, one should rather round up the number as a larger n means a narrower interval which is desirable. When deciding on a sample size, one must bare costs in mind, as the difference of a margin of error of 150 and one of 100 may differ in observation time and money quite a lot. Another thing to be taken into consideration is the fact that actual usable observations normally differ to the amount of observations planned beforehand due to, for example, non-response and drop-out rates.

There are a few conditions to using the formula $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$:

- The data should be from a SRS. Also note that undercoverage and nonresponse rates etc can create errors in the data.
- There is no bias involved when collecting the data.
- Try to search and correct outliers before using the formula. If correcting or removing the outliers is not possible, another formula less sensitive to outliers should be used.
- The population needs to be (relatively) Normal.
- The higher the sample size, the more accurate the formula.
- The population standard deviation σ is known.

6.2: Tests of Significance

In a significance test one accesses a certain hypothesis about observed data, this hypothesis being called a *Null hypothesis* (H_o). An example could be:
'H_o : There is no difference in the true means.'

There is an opposite statement to the null hypothesis, the alternative hypothesis (H_a). An example of this would be:

'H_a : The true means are not the same.'

Both of these are statements about the parameters of a population. H_a is the statement we hope to find evidence for, and H_o is the statement we hope to find evidence against. H_a can be one-sided, which states that the parameter differs from the null hypothesis in one specific direction, or it can be two-sided, stating that the parameter differs from the null hypothesis in both directions on a normal curve. Some statisticians argue that we should always use a two-sided alternative hypothesis. Whether the H_a is one-sided or two-sided must be decided on before one looks at the data. If one is not sure which to use, always use the two-sided alternative.When working with tests of significance we work from assuming that the H_o is true, and seeing how much evidence there is for it. The goal is to reject the null hypothesis, and so you can never accept it, only either reject it or realize there is not enough evidence to reject it. The H_a defines which directions count against the H_o.

To figure out how far away the parameter is from the null hypothesis, one needs to standardize the data estimate by using this formula: $z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$.

<u>P-Values</u>

The Supreme Court of the United States has stated that its benchmark for rejecting the H_o is two or three standard deviations, and this is used a lot, especially in the law. However, not all test statistics are normally distributed and therefore we rather use probability to express this. A test of significance has a *p-value*, which is the probability, assuming that the null hypothesis is true, of getting a value as extreme, or more extreme, than the observed value in this test. As we would like to reject the null hypothesis, a small as possible value of p is desirable, as it provides stronger evidence against it.

In order to find this p-value, one needs to look up the z-value found by using the formula $z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$ and looking up the corresponding probability in Table A of Introduction to the Practice of Statistics, 7th Ed (Moore, McCabe & Craig). Do not forget that with two-sided alternatives the probability of one side must be multiplied by two.

<u>Statistical Significance</u>

After figuring out the p-value, we need to conclude whether this is a small enough probability to use as sufficient evidence to reject the null hypothesis. This can be done by comparing the p-value with a fixed value regarded as the decisive value – this being called the *significance level α*. If the α is 0.05, this means that given the null hypothesis is true, the data will only give false evidence against the H_o 5% of the time. If the p-value is smaller, or as small as the significance level then we say that the data are statistically significant at that

level of α. This can be written as P<0.05, for example. Do however note that statistical significance does not necessarily mean that the data has practical significance.

These are the four steps to all tests of significance:

1. State the H_o and the H_a
2. Calculate the test statistic using the standardizing formula.
3. Calculate the corresponding p-value
4. Conclude significance or non-significance from the data. Keep this in mind when doing so:

- If the p-value>α -> The data does not provide sufficient evidence to reject the H_o.
- If the p-value < or = to α -> The H_o can be rejected and the alternative hypothesis is true.

Note: When a p-value is smaller than 0.001 then one reports it as P<0.001, as this value is sufficiently small enough to reject the null hypothesis!

Population mean Significance Testing

When using the significance test for the parameter population mean the null hypothesis is: H_o: the true population mean is μ_o. or H_o: $\mu = \mu_o$

In this case we use the formula: $z = \frac{\bar{x} - \mu o}{\sigma/\sqrt{n}}$ when finding the test statistic. The p-values can then be found according to:

- H_a: $\mu > \mu_o$ is $P(Z \geq z)$
- H_a: $\mu < \mu_o$ is $P(Z \leq z)$
- H_a: $\mu \neq \mu_o$ is $2P(Z \geq |z|)$

Two-sided significance tests and confidence intervals are comparable in this way:
At a significance level of α, the test rejects the null hypothesis for the population mean for values outside of the 1 – α confidence interval for μ.

The *critical value* of a standard normal curve is the value z* with a certain area to its right under that same curve.

Power in Inference testing

The *power* of a significance test to detect an alternative value than the one indicated in the null hypothesis is the probability that the significance test will reject the null hypothesis when the alternative is true. It is desirable to have a high power and 0.80 is the standard used by the US government agencies when using a significance level of 0.05. Below are the 3 steps to calculating the power:

1. State H_a, H_o, the alternative value and α.

2. Find all values of $\bar{x}$ with which we can reject the null hypothesis.
3. Find the probability of detecting these values when the alternative is true. The power is calculated with the formula: $P(Z = \frac{\bar{x} - \mu o}{\sigma/\sqrt{n}})$.

How to increase the power if it is too small:

- Increase α.
- Choose an alternative that is even further away from the value μo. The closer the values are to the μo, the harder they are to detect.
- Increase N.
- Decrease the standard deviation.

Note that failure to reject the null hypothesis when the power is low is not evidence that the null hypothesis can be accepted.

Two Types of Error

There are two types of error in statistical inference: Type I and Type II errors. A type I error happens when one rejects the null hypothesis when it is in fact true, where as a type II error is when one accepts the null hypothesis when in fact it is false. Type I error is given as α, as the significance level is the probability that a test will reject the null hypothesis when it is in fact true. The Type II error is given as β. See the table below for a better overview of how this all works.

	H_o is true	H_a is true
Reject H_o	Type I Error α	Correct (1- α)
Accept H_o	Correct (1-β)	Type II Error β

Note that (1- β) is the power of the significance test, so power against an alternative is 1 minus the Type II error for that alternative.

6.3 Use and Abuse of Tests

Nowadays, computer programs are commonly used to carry out significance tests; however, the use of a significance tests is not always easy:

- There is no boundary between significant and non-significant, instead you only have evidence that increases in strength as the p-value decreases.

- Sir R. A. Fisher, the man who invented formal statistical methods for analysing experimental data made this point: “A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.”
- A significance level α of 0.05 is the most commonly used.
- When samples are large, tiny deviations from the H_o become significant; however one needs to think about whether this statistical significance is really practical significance, by looking at the context of the experiment. Also, it is wise to state a confidence interval, and not only look at the significance tests.
- Be very aware of your data, and analyse it constantly, by looking out for outliers for example. Do not blindly type in the data information to a computer, always be critical.
- Take note that, on the other hand, even though there is no statistical significance, there may well be practical significance, for example certain small meaningful effects only detectable in large samples go unnoticed due to lack of statistical significance, even though these small effects are very practically significant.
- If a design of an experiment is flawed, no amount of statistical inference can correct that. This is due to the fact that the tests of significance and confidence intervals rely on the laws of probability, and are based on true randomization when sampling.
