Chapter 7. Distributional Statistical Inference

This chapter involves statistical inference for distributions and does not assume a known standard deviation any longer.

7.1: One-sample t distributions

One-sample t-tests

Instead of z-tests we will now look at t-tests and t-distributions. We will use the sample standard deviation *s* to estimate for the now unknown σ. When we do this (estimating the σ from data) we end up with a standard error, in this case, of the sample mean. This is calculated with this formula:

$$SE_{\bar{x}} = s / \sqrt{n}$$

When we substitute the standard error in the formula for a z-statistic you end up with the formula below for a one sample t statistic.

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

To test the null hypothesis: $H_o: \mu = \mu_o$ you use this formula to find the test statistic. The p-values can then be found according to:

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

Use Table D to find the p-values. This new statistic has a new distribution, called a *t distribution* and has (n – 1) *degrees of freedom*. We use t(k) to stand for the amount k of degrees of freedom in a t distribution. For an example of a t distribution see figure 7.1 in Introduction to the Practice of Statistics, 7th Ed (Moore, McCabe & Craig). Notice how the more degrees of freedom a t distribution has the more Normal it becomes, this is due to having a greater sample size. This also means that s will be more like σ the higher the n. It is also noticeable how the t-distribution has more probability in the extremes and less in the centre, as a result of the extra variability of now using an unknown population standard deviation.

One-Sample t confidence intervals

By filling in the standard error into the formula for *m* in the *z* confidence interval and by replacing *z** with *t** you get the formula for a t confidence interval:

$$\bar{x} \pm t^*(s/\sqrt{n})$$

where m = t*is the margin of error. This is accurate when the distribution is Normal and roughly accurate for a large sample size otherwise.

With confidence intervals, one normally gives the actual interval as an outcome of the formula, however sometimes it is preferred to report the mean and the margin of error instead.

Matched Pairs t procedures

Comparative studies are preferred over single-sample studies and there is one such comparative design that makes use of single-sample procedures: the *matched pairs* study. This is where participants are matched up in pairs according certain characteristics that are important to that study, and then their results are compared. Matched pairs are also used when randomization is impossible.

To do a matched pairs t procedure you use the differences between the two measurements as the data for the analysis and find the mean and the standard deviation of this data for the t test. The null hypothesis used focusses on if there is a difference or not. In most circumstances we cannot be sure of which direction to use in a one-sided test and so it is always safest to use a two-sided test, so $\mu = 0$ or $\mu \neq 0$.

Problems with matched pair procedures are that randomization is not possible, data may not be Normal and sample sizes can be very small.

T-test robustness

It never occurs that real populations are exactly Normal, and t procedures depend on Normality. Robustness is how easily affected by non-Normality a procedure is. The less strongly affected the more robust. This applies to other violations of the assumptions made about statistical inference.

T procedures are fairly robust against non-Normality except for when there is strong skewness or outliers involved. T procedures are not robust against outliers as $\bar{x}$ and *s* are not resistant to outliers. If $n < 15$ you can use the t test if the data is close to Normal only, and no outliers are present. If $n \geq 15$ then you can use the t test always except for when there are outliers or strong skewness. When $n \geq 40$ the t test can always be used, even in the case of skewness.

Power of the t-test

Power in statistical testing is its capability to identify deviations from H_o, and so high power is desirable as we always want to prove the null hypothesis false. As noted in the previous chapter: 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. Calculate the power in a t test by:

1. Choose the standard deviation, the alternative value, α and whether the test is one- or two-sided.
   Note: it is always better to use a s that is slightly bigger than the one we expect than to use a smaller one. Use this when rounding up values for the standard deviation.

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.

Non-Normal populations

All inference tests taught so far is based on Normality. Here are 3 tactics to deal with non-Normality, due to a small sample size:

- If the distribution can be described by a different standard distribution, then use inference procedures based on that model.
- If the problem is skewness, one can transform the data before performing t tests and confidence intervals. The most used transformation is logarithm, which pull the right tail of the distribution in. So instead of analysing the original values analyse the logarithms of these values, which are much less skewed.
- Some inference procedures do not need a particular population distribution. You can therefore use these so-called *nonparametric* or *distribution-free procedures*. This tactic does have its downsides though; these tests are less powerful than the t test and they focus on the median and therefore do not ask the same questions as a t test would.

A *sign test* is the simplest and most useful nonparametric procedure. All the variations of the sign test are based on counts and the binomial $B(n,½)$ distribution. We will only discuss the sign test for matched pairs here. The null hypothesis in this case is: H_o: $p = ½$ and the alternative can be both one- and two-sided, depending on the case. Recall the knowledge you already have about matched pairs, and remember this looks at the differences between the pairs. When doing a sign test for matched pairs, ignore the pairs with no difference and *n* becomes the number of left-over pairs. The test statistic is now the count X of all the pairs with a positive difference. Do not forget that this test is testing the H_o that the median, not the mean, of the differences is 0.

7.2: Two-sample Statistical Inference

In two-sample problems the aim is to infer by comparing two independent and randomized groups from the same population. Each group is exposed to different treatments and the two groups may differ in size. Comparing two randomized groups from different populations can also be a two-sample problem. There is no matching up of pairs in two-sample problems. You tell the two groups apart by giving each group a number and putting the number as an subscript next to each statistic for that group, for example μ_1, μ_2, n_1, n_2, σ_1 and σ_2. We test the null hypothesis for no difference: H_o: $\mu_1 = \mu_2$. An overview of the notations used in the different groups is given here below:

Population	Variable	Mean	Standard Deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

Sample Group	Sample Size	Sample Mean	Sample Standard Deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

Two-sample z tests

To compute the difference in means one uses $\bar{x}_1 - \bar{x}_2$. This is Normally distributed if the two population distributions are Normal.To compute the difference in variance one uses the addition rule for variances:

.

The two-sample statistic is now:

$$z = \frac{(\bar{x}1 - \bar{x}2) - (\mu 1 - \mu 2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

This has a Normal N(0,1) sampling distribution and the groups were randomly sampled. Note this is for the unlikely event that both standard deviations are known.

i. *Two-sample t procedures*

Substituting the σ's with the known *s*'s in the z-test formula gives the two sample t statistic:

$$t = \frac{(\bar{x}1 - \bar{x}2) - (\mu 1 - \mu 2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Unfortunately this does not have a t distribution as we have replaced not just a single standard deviation by its sample standard deviation but two instead and this no longer gives such a distribution. However we are able to approximate the distribution by using an approximation of the degrees of freedom k. You can either do this by using software or by choosing the k that is smaller one of $n_1 - 1$ and $n_2 - 1$.

- *Significance test:* The two sample t significance test formula, to test the null hypothesis H_o: $\mu_1 = \mu_2$, is:

$$t = \frac{(\bar{x}1 - \bar{x}2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where the degrees of freedom are either approximated by software or are the smaller of n_1 – 1 and n_2 – 1.

Confidence Interval: The same ideas count for the two-sample t confidence interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*.$$

The degrees of freedom are calculated in the same way as with the significance test.

Two-sample robustness

Two-sample t procedures are even more robust than one-sample t procedures, and are the most robust against non-Normality when sample sizes of both groups are the same. Because of this it is advisable to have the same sample sizes for both groups wherever possible.

When choosing which population to label 1 and which 2, it is best to choose the one with the higher statistic as 1 so as not to get a negative value for t.

Statistical inference for small samples

Often you can draw conclusions about a population even when the sample size is small. Especially with a big effect size as this should then be noticeable even with smaller n's.

Pooled t procedures

When two Normal population distributions have the same σ^2 then you get exactly a t distribution. To do this you must find the pooled estimator of σ^2:

$$s_p^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) / (n_1 + n_2 - 2).$$

now becomes: $\sigma^2 (\frac{1}{n1} + \frac{1}{n2})$.

The z test therefore becomes:

$$z = \frac{(\bar{x}1 - \bar{x}2) - (\mu1 - \mu2)}{\sigma 2(\frac{1}{n1} + \frac{1}{n2})}$$

From all of this we can derive the *pooled two-sample t inference procedure*:

$$t = \frac{(\bar{x}1 - \bar{x}2)}{\text{sp}2(\frac{1}{n1} + \frac{1}{n2})}$$

The degrees of freedom in this case are $n_1 + n_2 - 2$. The p-values can then be found according to:

- H_a: $\mu_1 > \mu_2$ is $P(T \geq t)$
- H_a: $\mu_1 < \mu_2$ is $P(T \leq t)$
- H_a: $\mu_1 \neq \mu_2$ is $2P(T \geq |t|)$

A C level confidence interval for $\mu_1 - \mu_2$ is calculated using:

$$(\bar{X}_1 - \bar{X}_2) \pm t^* s_p^2 (\frac{1}{n1} + \frac{1}{n2}).$$

This pooled two-sample t inference procedure requires the assumption that both standard deviations are the same, which is very difficult to confirm, making this procedure quite risky to use. This procedure is quite robust against non-Normality and unequal standard deviations in the cases of the sample sizes being nearly the same. Unless the n's are large, use the pooled two-sample t inference procedure with caution when the n's are different in size and there are unequal standard deviations.

7.3 Statistical Inference for variance

Not only can you do inference for means, but also for spread. Unlike the t procedures for means, the inference tests for standard deviations is not robust at all against non-Normality, but instead very sensitive, even in large samples. It is best not to use these tests in basic statistical practice. Due to the limited usefulness in these procedures, we will only discuss one; the F test.

Suppose we have H_o: $\sigma_1 = \sigma_2$ and Ha: $\sigma_1 \neq \sigma_2$, then the F statistic is:

$$F = s_1^2 / s_2^2$$

The samples must be SRS's and be drawn from Normal populations. The degrees of freedom are n_1-1 and n_2-1 when H_o is true. The numerator degrees of freedom are always mentioned first. The order is important as doing it the other way around changes the F distribution. The notation for this distribution is *F(j,k)* with *j* degrees of freedom in the numerator and *k* of the denominator. A F distribution is not symmetric like a Normal distribution but instead is right skewed. See Figure 7.17 in Introduction to the Practice of Statistics, 7th Ed (Moore, McCabe & Craig) for an example of this distribution. A F statistic cannot take a negative value as standard deviations can only be positive, and the probability can never be below 0. The peak of the distribution is close to 1 and the further the value from 1 in both directions the more evidence against the H_o. Use table E to find the critical values for these F distributions.

If you are not using statistical software arrange the F test as: F = larger s^2 / smaller s^2. That way you cannot get a F smaller than 1. Compare F with the critical values in Table E and double the p-value for two-sided F tests. Remember the order of the degrees of freedom is important, also when using Table E.
The robustness of the one- and two-sample t tests is very remarkable, but the lack of robustness in the variance tests is as remarkable. It is best not to use these tests at all.

<u>Power of two-sample t tests</u>
To find the power of the pooled two-sample t test, considering only the case where H_o: $\mu_1 = \mu_2$, state the alternative value, the sample sizes, α, a guess of σ, the degrees of freedom ($df=n_1+n_2-2$) and the value of t* that will lead to rejection of the H_o. You also need to calculate the *noncentrality parameter* for the alternative of interest using:

$$\delta = \frac{|\mu 1 - \mu 2|}{\sigma\left(\frac{1}{n1} + \frac{1}{n2}\right)}.$$

You can perform this calculation by using software or by approximating the power as $P(z > t^* - \delta)$ and by using table A. The denominator in this formula is the estimate of the standard error. This can be used in the margin of error formula when calculating confidence intervals.
If however we do not assume both standard deviations are equal then we use the formula below as the standard error estimate and in the denominator of the previous formula. All of this gives a *noncentral t distribution*.

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$