9231 CAIE Further Maths Stats — Inferential Statistics

Alston

1 t-distribution

Used to model data (variance) about a population when the sample size is small (when the sample is not big enough to use normal). Generally it's when $n < 30$.

To use the t-distribution, you must have:

- Underlying normal distribution assumed
- Variance of parent distribution is unknown
- Small sample size

The unbiased estimator is the same:

Definition 1.1 – Unbiased Estimator

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) = \frac{1}{n-1}\left(\sum x^2 - n\bar{x}^2\right)$$

And we also have the test stat:

Definition 1.2 – Test Statistic

$$t = \frac{x-\mu}{\frac{s}{\sqrt{n}}}$$

Note that here in the denominator we are dividing by the **standard error of the mean**. This comes from $Var(\bar{X}) = Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{1}{n^2}Var(X_1+X_2+\cdots+X_n) = \frac{1}{n^2}\times n \times Var(X) = \frac{s^2}{n}$.

And then we are just dividing by $\sqrt{Var(\bar{X})}$.

2 Difference in Means

Here, we have assumptions

- Two independent distributions
- Same population variance across them

- Normal underlying distributions
- Each sample size are sufficiently large enough

So when we are testing for the difference in means, we are just combining the two sample mean distributions. So if we take $\bar{X} - \bar{Y}$, their mean will be the difference, and the variance of the resultant distribution will be the sum of the variences of $\bar{X}$ and $\bar{Y}$.

Definition 2.1 – Z Value

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

Note that here, often we have $\mu_x - \mu_y = 0$ as a result of the H_0. And note here that we use sample variation here, but the theoretical test statistic would have σ_x^2 and σ_y^2.

2.1 Small sample sizes

If $n < 15$, we would need to pool our variances into one variance.

So we are literally just taking all the data from both samples, and treat as if they are from one sample when calculating the variance.

But furthermore, if you know the unbiased estimator of the variance for both samples you can change up the formula. Originally it's this:

$$s_p^2 = \frac{\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2}{n_x + n_y - 2}$$

But if we know $s_x^2 = \frac{\sum(x-\bar{x})^2}{n_x - 1}$ and likewise for y, then we have

$$s_p^2 = \frac{(n_x - 1)s_x^2 - (n_y - 1)s_y^2}{n_x + n_y - 2}$$

And so unlike in section 2, we will use t-distribution here due to a small sample size. Remember to change up your degree of freedom with how many parameters you've estimated! If we make the assumption that the underlying variances for the two population samples are the same, we can change the denominator of the test statistic to $\sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$

3 Paired t-tests

A paired t-test looks for the changes in each entry before and after an action is performed on it.

Definition 3.1 – Test Statistic

$$t = \frac{\bar{d} - k}{\frac{s_d}{\sqrt{n}}}$$

$\bar{d}$ is just the mean of the differences. H_0 assumes that the mean is k. Normally this would be equal to 0. Finally, the bottom is the standard deviation of the distribution (normal) of the differences.

4 Confidence Intervals

This can be thought of as just the acceptance region of the test. We assess whether the CI actually contains the population mean. Here, if n is small, we will use the t-distribution.

Definition 4.1 – Confidence Intervals

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$$

The $\frac{s}{\sqrt{n}}$ term is actually just the standard error of the mean. Then the t value is just the value read from the table. the $\frac{\alpha}{2}$ is the percentage of the confidence interval. So if we want a 90% confidence interval, that leaves 5% on both sides, so we would want the t value at 95%.

5 Confidence Interval for Difference in Means

This is just part 2, converted to a confidence interval. We need the assumption that n is large enough $n \geq 30$, so here we will use the z score.

Definition 5.1 – Confidence interval, difference in means

$$\bar{x} - \bar{y} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

5.1 Pooled Variance

If $n < 30$, we would need to pool the data from both samples for the variance.

Definition 5.2 – CI for small samples

$$(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2},n_x+n_y-2} \times s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$