§1 Hand-out 1

Dr. M. P. M. M. McLoughlin

Version 2 Revised 2018

1. Fundamentals

1.1. **Preliminaries.** Let $X \sim f(x)$, $f : \mathbb{R} \longrightarrow \mathbb{R}$, such that f is a well-defined probability density function and let $Y \sim g(y)$, $g : \mathbb{R} \longrightarrow \mathbb{R}$, such that g is a well-defined probability density function.

The population parameters for X are the mean, μ_X , the variance, σ_X^2, the standard deviation, σ_X, etc. and the population parameters for Y are the mean, μ_Y, the variance, σ_Y^2,the standard deviation, σ_Y, etc.

Let X and Y have a well-defined joint probability density function, $h((x, y))$ where $h : \mathbb{R}\times\mathbb{R} \longrightarrow \mathbb{R}$. Hence, the population covariance is σ_{XY} formed by computing $E[XY]-E[X]\cdot E[Y]$ and the population (Pearson product-moment) correlation between X and Y is $\rho_{XY} = \dfrac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$ such that $\sigma_X \neq 0$ and $\sigma_Y \neq 0$.

1.2. **Sample Statistics.** Let $D = \{X_1, X_2, X_3, ..., X_n\}$ be a finite data set from the X-population of interest such that $n \in \mathbb{N}$ and
$C = \{Y_1, Y_2, Y_3, ..., Y_m\}$ be a finite data set from a Y-population of interest such that $m \in \mathbb{N}$.

We assume such samples were random; i.i.d.; per variable there was no error in measurement or recording of the data.

Recall the population parameters for X are the mean, μ_X, the standard deviation, σ_X, etc.; $\overline{X}$ is an estimator for μ_X, $\widehat{\mu_X}$; S_X^2 is an estimator for σ_X^2, $\widehat{\sigma_X^2}$; and S_X is an estimator for σ_X, $\widehat{\sigma_X}$. Please recall when any of these estimators are unbiased, consistent, or sufficient and how they were created (either with the method of moments or the method of maximum likelihood).

Likewise the population parameters for Y are the mean, μ_Y, the standard deviation, σ_Y, etc.; $\overline{Y}$ is an estimator for μ_Y, $\widehat{\mu_Y}$; S_Y^2 is an estimator for σ_Y^2, $\widehat{\sigma_Y^2}$; and S_Y is an estimator for σ_Y, $\widehat{\sigma_Y}$.

The sample (Pearson product-moment) correlation is between X and Y is r_{XY} (only in the case equal sample sizes so $m = n$). We use $r_{XY} = \widehat{\rho_{XY}}$.

Recall also:

$$\overline{X} = \sum_{j=1}^{n} \frac{X_j}{n} \qquad \overline{Y} = \sum_{k=1}^{m} \frac{Y_k}{m}$$

$$S_X = \sqrt{\frac{\sum_{j=1}^{n}(X_j - \overline{X})^2}{(n-1)}} \qquad S_Y = \sqrt{\frac{\sum_{k=1}^{m}(Y_k - \overline{Y})^2}{(m-1)}}$$

$$S_{\overline{X}} = \frac{S_X}{\sqrt{n}} \qquad S_{\overline{Y}} = \frac{S_Y}{\sqrt{m}}$$

$$r_{XY} = \frac{\sum_{a=1}^{n}((X_a - \bar{X}) \cdot (Y_a - \bar{Y}))}{\sqrt{(\sum_{a=1}^{n}(X_a - \bar{X})^2) \cdot (\sum_{a=1}^{n}(Y_a - \bar{Y})^2)}}$$

2. Student T-statistic

2.1. **Independent t-test One Sample Statistical Inference About the Mean of a Population.** Set α. Decide on n. State the hypothesis (the population parameter being estimated (1) is some real number, m; (2) is less than or equal to some real number, m; or, (3) is greater than or equal to some real number, m.).

If an assumed σ_X is used, then the test is a Z-test. Let us proceed with the scenario that there is not an assumed σ_X, so the statistic is built as a T-statistic with $df = n - 1$.

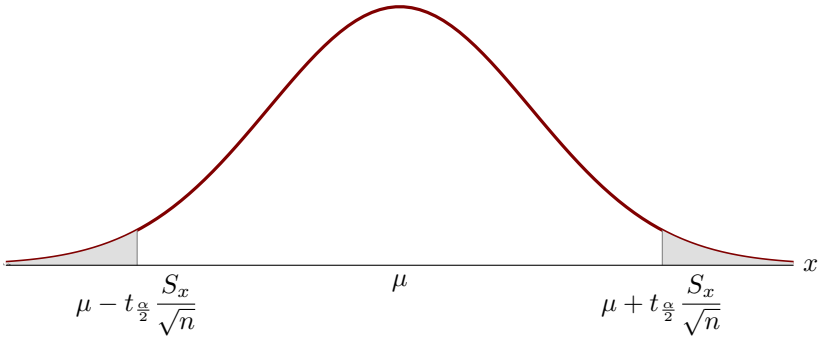
So, in case (1) we have:
$H_0 : \mu = m$ versus the alternate hypothesis $H_A : \mu \neq m$.

Case (2) is:
$H_0 : \mu \leq m$ versus the alternate hypothesis $H_A : \mu > m$.

Whilst case (3) is:
$H_0 : \mu \geq m$ versus the alternate hypothesis $H_A : \mu < m$.

Data is collected and the sample is assumed random (pseudo-random); the observations were i.i.d.; and, there was no error in measurement or recording of the data.

The distribution of the $\bar{X}$ is either normal (where $X \sim N(x, \mu_X, \sigma_X)$) or via appealing to the Central Limit Theorem $\bar{X}$ is approximately normal. So, in either case (where $\bar{X} \sim N(x, \mu_{\bar{X}}, \sigma_{\bar{X}})$) or $\bar{X} \dot{\sim} N(x, \mu_{\bar{X}}, \sigma_{\bar{X}})$). The centre of the distribution under H_0 is the $\mu = m$. Let $t_{\alpha/2}$ be the t-critical value for T-statistic with $df = n - 1$ and α-level set (as was done apriori). So, we have:



Distribution under H_0

Let $T_{comp} = \dfrac{\bar{X} - \mu_{\bar{X}}}{S_{\bar{X}}}$. When T_{comp} falls in the shaded 'regions,' then there is sufficient evidence at the α-level to reject the H_0. When T_{comp} falls in the white regions (between the shaded 'regions'), then there is not sufficient evidence at the α-level to reject the H_0 (fail to reject the null hypothesis).

Now consider that the $\bar{X}$ is a point-estimate for $\mu_{\bar{X}}$ based on the sample; might there not be a better way to report the results from the study? The answer is a qualified, 'yes.' Such is called a "confidence interval." It is constructed as follows. Recall that the α- level was *a priori* set. So, $(\bar{X} - t_{\frac{\alpha}{2}} \dfrac{S_x}{\sqrt{n}}, \bar{X} - t_{\frac{\alpha}{2}} \dfrac{S_x}{\sqrt{n}})$ is a segment with $\bar{X}$ as it centre having the property that it generalises the point estimate (this segment based on the data varies from sample-to-sample) and if said segment happens to 'cross' $\mu_{\bar{X}}$ (which is fixed) then there is sufficient evidence at the α-level to reject the H_0. However, if said segment does not 'cross' $\mu_{\bar{X}}$ (which is fixed), then there is NOT sufficient evidence at the α-level to reject the H_0.

2.2. **'Small' (not large) Sample Size Independent t-test.** Independent t-test Pooled Variance Formula with variances assumed equal with *df* being $n + m - 2$

$$S_{pool-equal} = S_p = \sqrt{\frac{(n-1)\cdot S_X^2 + (m-1)\cdot S_Y^2}{n+m-2}}$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)\cdot S_X^2 + (m-1)\cdot S_Y^2}{n+m-2} \cdot (\frac{1}{n} + \frac{1}{m})}}$$

which is:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

2.3. **'Large' Sample Size Independent t-test.** Independent Pooled Variance Formula with variances assumed equal for 'large' sample sizes $m \geq 30, n \geq 30$ with *df* being $n + m - 2$

$$S_{pool-equal} = S_p = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

2.4. **Independent t-test Non-equal Variances.** Independent Pooled Variance Formula with variances are not assumed equal. Large sample:

$$S_{pool-non-equal-large} = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

Small sample:

$$S_{pool-non-equal-small} = \sqrt{\frac{(n-1)\cdot S_X^2 + (m-1)\cdot S_Y^2}{n+m-2}}$$

Then,

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)\cdot S_X^2 + (m-1)\cdot S_Y^2}{n+m-2} \cdot (\frac{1}{n} + \frac{1}{m})}}$$

with *df* being adjusted

$$df \approx \frac{(\frac{S_X^2}{n} + \frac{S_Y^2}{m})^2}{\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1}}$$

We shan't use this and I recall that when I was in school we were strongly encouraged NOT to ever do an independent t-test with non-equal of variances.

2.5. **Paired Sample t-test.** If one is doing this by hand, first compute r_{xy}; recall, $r_{XY} = \frac{\sum_{a=1}^{n}((X_a-\bar{X})\cdot(Y_a-\bar{Y}))}{\sqrt{(\sum_{a=1}^{n}(X_a-\bar{X})^2)\cdot(\sum_{a=1}^{n}(Y_a-\bar{Y})^2)}}$
since the samples are related (two measures from the same subject or matched pairs), the correlated data formula is used.

Clearly (a dangerous word) the formulae are already noted except the T-statistic:

$$t_{paired} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_X^2}{n}\right) + \left(\frac{S_Y^2}{n}\right) - 2 \cdot r_{XY} \cdot (\frac{S_X \cdot S_Y}{n})}}$$

with *df* number of pairs minus one.

All the hypothesis testing verbiage stays the same.

2.6. **Programming.** Minitab, SPSS, R, SAS, etc.

2.7. **Exercises.**

Exercise 2.1. *A study is to be conducted to compare the weights of cats and dogs. Let us assume the population(s) are normally distributed with μ_{cat}; μ_{dog}; σ_{cat}; and σ_{dog} all existing such that $\sigma_{cat} = \sigma_{dog}$ and $\sigma_{cat} \neq 0$.*

Set your α level. Determine what your practical null hypothesis is; and, based on that determine what your statistical null hypothesis versus alternate hypothesis is.

What statistical test(s) is (are) to be done?

Assume you are able to take a pseudo-random sample of cats and dogs of size 5 for the cats and 4 for the dogs (one of the dogs got let out (whoop whoop) so we have but 4). Assume the observations are independent identically distributed (i.i.d.) and there is no error of measurement in the collection or recording of the data.

Weights of cats: 21, 35, 13, 21, 10. Weights of dogs: 31, 10, 20, 40.

Do the work completely by hand. What are the results (properly stated)? Do you opine, based on the samples, there is any difference between the weights of cats and dogs?