Lecture 2: Welch test vs Mann-Whitney test

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

November 13, 2023

Plan

Plan
1) Welch test vs t-test
2) Non-parametric statistics
3) Mann-Whitney U test

Plan

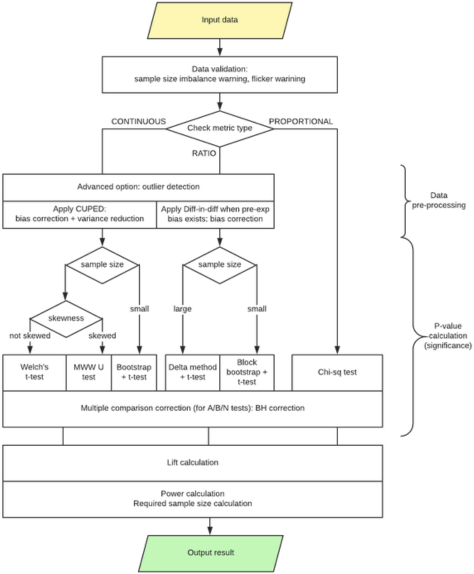


Figure: Uber's statistics engine is used for A/B/N experiments

Welch test – unequal variances
The two-sample t-test for unpaired data is defined as:

$$\begin{array}{ll} \mathrm{H}_0: & \mu_1 = \mu_2 \\ \mathrm{H}_a: & \mu_1 \neq \mu_2 \\ \text{Test Statistic:} & T = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \end{array} \tag{1}$$

where N_1 and N_2 are the sample sizes, $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means, and s_1^2 and s_2^2 are the sample variances.
Critical Region:
Reject the null hypothesis that the two means are equal if

$$|T| > t_{1-\alpha/2,\nu}$$

where $t_{1-\alpha/2,\nu}$ is the critical value of the t distribution with ν degrees of freedom where

$$v = \frac{\left(s_1^2/N_1 + s_2^2/N_2\right)^2}{\left(s_1^2/N_1\right)^2/(N_1-1) + \left(s_2^2/N_2\right)^2/(N_2-1)}$$

t-test – equal variances
If equal variances are assumed, then the formula reduces to:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$

where

$$s_p^2 = \frac{(N_1-1)\,s_1^2 + (N_2-1)\,s_2^2}{N_1 + N_2 - 2}$$

If equal variances are assumed, then $\nu = N_1 + N_2 - 2$

Two-sample t-test vs Welch test

Consider the setting of two independent samples $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1, \ldots, Y_m \overset{IID}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$, as in Example 7.2 of last lecture. Here μ_X, μ_Y, σ^2 are all unknown; note that we are assuming (for now) a common variance σ^2 for both samples. For the testing problem

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X > \mu_Y$$

a natural idea is to reject H_0 for large values of $\bar{X} - \bar{Y}$. Observe that $\bar{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma^2}{n}\right)$, $-\bar{Y} \sim \mathcal{N}\left(-\mu_Y, \frac{\sigma^2}{m}\right)$, and these are independent. Then their sum is distributed [1] as

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

Two-sample t-test vs Welch test

Under $H_0, \mu_X - \mu_Y = 0$, so $(\bar{X} - \bar{Y})/\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}} \sim \mathcal{N}(0,1)$. If σ^2 were known, then a level- α test based on $\bar{X} - \bar{Y}$ would reject when

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} > z(\alpha)$$

Since σ^2 is unknown, we estimate it from the data. We may use both the X_i 's and Y_i 's to estimate σ^2 by taking the pooled sample variance

$$S_p^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right),$$

and take as a test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Two-sample t-test vs Welch test

To derive the null distribution and rejection threshold for T, we may rewrite this as

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} / \sqrt{S_p^2/\sigma^2}$$

By independence of the two samples, $\bar{X}, \bar{Y}, \sum_i (X_i - \bar{X})^2, \sum_j \left(Y_j - \bar{Y}\right)^2$ are all independent, with the last two quantities distributed as $\sigma^2\chi^2_{n-1}$ and $\sigma^2\chi^2_{m-1}$. Then under H_0,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0,1), \quad \frac{S_p^2}{\sigma^2} \sim \frac{1}{m+n-2}\chi^2_{m+n-2},$$

and these are independent. So the distribution of T is the same for all data distributions $P \in H_0$ and is given by

$$T \sim t_{m+n-2}.$$

The test that rejects H_0 when $T > t_{m+n-2}(\alpha)$ (the upper α point of the t_{m+n-2} distribution) is called the two-sample $\boldsymbol{t}$-test.

Two-sample t-test vs Welch test

The problem is that we don't have a nice description of what the difference between two (scaled) t distributions looks like. For normal distributions, we can use the very convenient fact that the sum or difference of normal random variables is also normal; that is not the case for t distributions!
In principle, because we know the probability density functions of T_{n_A-1} and T_{n_B-1}, and they are independent, we could calculate the probability density function of the random variable listed above for particular values of the parameters.
But that does not solve our problem, because we need to write down a distribution that is independent of the test parameters (i.e., that does not depend on S_A and S_B).

Two-sample t-test vs Welch test

In most situations when we are comparing two populations, it is not reasonable to assume that the population variances are the same. For this reason, various unpooled two-sample t tests have been developed.
The most popular such test is known as Welch's unequal-variances t test.
Welch's t-test is generally more accurate than Student's equal-variances t test (described above) in the situation where the two sample variances are far apart, or when the sample sizes differ drastically.

Two-sample t-test vs Welch test

The assumption of common variance σ^2 for the two samples is oftentimes problematic (and violated) in practice. If we assume instead that $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \overset{IID}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ for possibly different values of σ_X^2 and σ_Y^2, then $\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{1}{n}\sigma_X^2 + \frac{1}{m}\sigma_Y^2$, and we may estimate this by $\frac{1}{n}S_X^2 + \frac{1}{m}S_Y^2$, where S_X^2 and S_Y^2 are the sample variances of the two samples. Then we may use the test statistic

$$T_{\text{welch}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n}S_X^2 + \frac{1}{m}S_Y^2}}.$$

The distribution of T_{welch} under H_0 is no longer exactly a t distribution, but it was shown by Welch (1947) to be close to the t distribution with

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

degrees of freedom. The test that rejects when T_{Welch} exceeds the upper α point of this t distribution is called Welch's ***t***-test or the unequal variances ***t***-test.

Note:

Textbooks frequently recommend the calculation of number of degrees of freedom, rounded down to the nearest integer. However, rounding down tends to produce a conservative test.

More generally, some text-books recommend rounding to the nearest integer.

Welch's result is quite technical, but we can describe roughly where the formula for the degrees of freedom comes from.
The idea is to rewrite the quotient in the test statistic (in the same way we did in the theorem earlier) and try to write the denominator ratio as the sum of squares of independent standard normals.
This cannot be done exactly, but if it could, we would then be able to find the number of terms by using the method of moments to compare the means and variances of the two expressions.
Carefully going through the calculations eventually yields the degree-of-freedom formula given on the previous slide.

Two-sample t-test vs Welch test

We make a few brief remarks about when to use these various t tests.

- Most sources still identify Student's t test as the preferred test to use when the sample variances are not far away from each other, and give various approximate rules for deciding what "far away" means (e.g., requiring the variances not to differ by a factor of more than 2).
- When the sample variances are far apart, Welch's t test tends to give more reliable results (in the sense of having lower type I and type II error probabilities). Even when the sample variances are close, Welch's t test is generally not that much worse than Student's t test (which has a higher power in the situations where it should be used).
- Neither test is exact (in the sense that it gives exact p-values) except in the case of Student's t test where the population variances are equal. In practice, this means that the type I error rate will deviate somewhat from the desired significance level α.
- Welch's t test tends to maintain a type I error rate closer to the desired significance level α than Student's t test does (although of course there are scenarios in which it is worse).
- It is also worth noting that, as the sample sizes of both groups become large, both tests are very close to the two-sample z test we have previously described. In practice, with samples larger than 100-200 or so, there is a negligible difference between the results of these t tests and the simpler z test

Two-sample t-test vs Welch test

- Welch's t-test is more robust than Student's t-test and maintains type I error rates close to nominal for unequal variances and for unequal sample sizes under normality.
- Furthermore, the power of Welch's t-test comes close to that of Student's t-test, even when the population variances are equal and sample sizes are balanced.
- Welch's t-test can be generalized to more than 2-samples, which is more robust than one-way analysis of variance (ANOVA).
- It is not recommended to pre-test for equal variances and then choose between Student's t-test or Welch's t-test. Rather, Welch's t-test can be applied directly and without any substantial disadvantages to Student's t-test as noted above.
- Welch's t-test remains robust for skewed distributions and large sample sizes.

Non-parametric statistics

Any statistical analysis, including those outside of the scope of this course, involves the combination of:

- information that comes from assumptions, and
- information that comes from data

Assumptions that we typically have to contend with include those regarding:

- representativeness of the sample
- accuracy of the data
- underlying relationship(s) between key covariates

One way of classifying statistical analysis methods is on the basis of the extent of parametric assumptions regarding T, and how it depends on $\boldsymbol{X}$ or, more formally, regarding the conditional distribution of $T \mid \boldsymbol{X}$
For example, consider a linear regression analysis for a continuous response, T, based on the following components/assumptions:

1. mean model: $\mathrm{E}\left[T_i \mid \boldsymbol{X}_i\right] = \boldsymbol{X}_i^T \boldsymbol{\beta}$
2. error term: $\epsilon_i = T_i - \mathrm{E}\left[T_i \mid \boldsymbol{X}_i\right]$
3. the ϵ_i 's are independent
4. $\epsilon_i \sim$ Normal $\left(0, \sigma^2\right)$

Implicitly, this specification corresponds to assuming:

$$T_i \mid \boldsymbol{X}_i \sim \text{Normal}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}, \sigma^2\right)$$

This is a fully parametric analysis because it characterizes the entire distribution of $T \mid \boldsymbol{X}$

Now suppose we relax the 4th component of the previous specification as follows:

1. mean model: $\mathrm{E}[T_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i^T \boldsymbol{\beta}$
2. error term: $\epsilon_i = T_i - \mathrm{E}[T_i \mid \boldsymbol{X}_i]$
3. the ϵ_i 's are independent
4. $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$

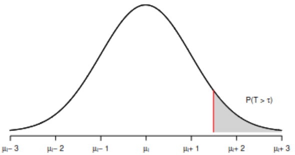
From this, all we know about the distribution of the outcome is that:

$$\mathrm{E}[T_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i^T \boldsymbol{\beta}$$
$$\mathrm{Var}[T_i \mid \boldsymbol{X}_i] = \sigma^2$$

Thus, this model is semi-parametric $\star$ there is some structure, specifically in how the mean and variance vary (or not) across levels of $\boldsymbol{X}$ $\star$ structure does not give us the entire distribution

Non-parametric statistics

One benefit of specifying a fully parametric analysis is that we then can use a likelihood-based analysis such as maximum likelihood
$\star$ if the model specification is correct, such an analysis is the 'best'
A second benefit is that, following estimation, one can use the results to calculate a whole range of potentially interesting quantities
$\star$ e.g. the probability that, for a given covariate profile, the outcome is greater than some (clinically-relevant) threshold, say τ



If, however, it is not the case that all of the assumptions that underpin the parametric specification hold and yet we assume it to be the case, then we are not guaranteed to get valid results

- potential for bias
- potential for incorrect inference

For example, if we assume $\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$ in a linear regression analysis but this is not the case then standard errors, and hence confidence intervals and p-values, will, in general, be biased

If we are uncertain, it may be 'safest' to proceed with the semi-parametric model

$\star$ although you still have to contend with whether $\mathrm{E}\left[\epsilon_i\right] = 0$ and $\text{Var}\left[\epsilon_i\right] = \sigma^2$ truly hold

If this is the path that is pursued, it's important to note that there are some potential drawbacks

Non-parametric statistics

Moreover, because the specification does not tell us everything about the distribution we are limited in what we can (easily) estimate
This may not be a big deal, however, if we only care about $\boldsymbol{\beta}$
Another consequence, however, is that we cannot use a likelihood-based analysis
- $\star$ we need other statistical tools
- $\star$ e.g. in the case of the linear regression analysis we use least squares estimation

Non-parametric analyses
A non-parametric statistical analysis procedure places no structure on the distribution of T
$\star$ in terms of the shape as well as how differences across levels of $\boldsymbol{X}$ manifest
In general, a non-parametric statistical analysis procedure in the context of this course places no structure on the distribution of the response variable
$\star$ i.e. on $T_i \mid \boldsymbol{X}_i$

Example 1 (Non-parametric density estimation – Histogram):

$$p(x) = 1/2p_{(0,1)} + 1/10 \sum_{j=0}^{4} p_{(j/2-1,1/10)}(x)$$

$p_{(}\mu, \sigma)$ - density of $N(\mu, \sigma)$
$p_1, \dots, p_k$
$a_1 + \dots + a_k = 1, a_i \geq 0$
$\sum_{i=j}^{k} a_i p_i(x)$ j- probability density (mixtures)

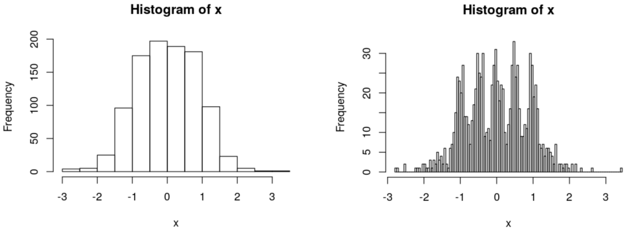


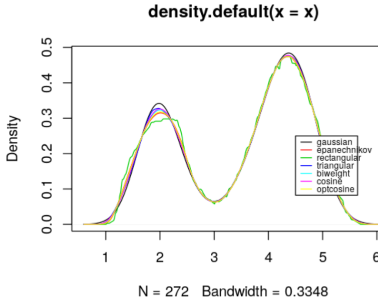
Figure: Breaks 'Scott' vs 'Freedman-Diaconis'

Non-parametric statistics

Example 2 (Non-parametric density estimation – KDE):
Let $(x_1, x_2, \dots, x_n)$ be independent and identically distributed samples drawn from some univariate distribution with an unknown density f at any given point x. We are interested in estimating the shape of this function f. Its kernel density estimator is

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

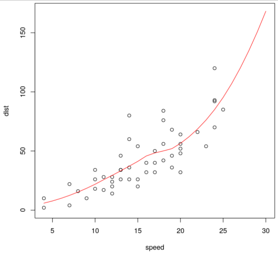
where K is the kernel - a non-negative function - and $h > 0$ is a smoothing parameter called the bandwidth. A kernel with subscript h is called the scaled kernel and defined as $K_h(x) = 1/hK(x/h)$. Intuitively one wants to choose h as small as the data will allow; however, there is always a trade-off between the bias of the estimator and its variance. The choice of bandwidth is discussed in more detail below.



Example 3 – LOESS (local regression) = LOWESS (locally weighted scatter plot smoothing)
Clevland (1979):
at the vicinity of x_i : $f(x) = \beta_0^{(i)} + \beta_1^{(i)} x$
$\sum_{j=1}^{n} w_i(x_j)(Y_j - \beta_0^{(i)} + \beta_1^{(i)} x)^2$
Tricube function

$$K(x) = \begin{cases} (1 - |x|^3)^3, |x| < 1 \\ 0, |x| \geq 1 \end{cases}$$

$w_i(X_j) = 1/hK(\dfrac{X_j - X_i}{h})$, h - bandwidth
Algorithm:
1. $w_i(X_j)$ 2. (*) $\rightarrow \beta_0^{(i)}, \beta_1^{(i)}$ 3. $e_i = Y_j - \beta_0^{(i)} + \beta_1^{(i)} X_i$ 4. $w_i(X_j) \rightarrow \delta_j w_i(X_j)$

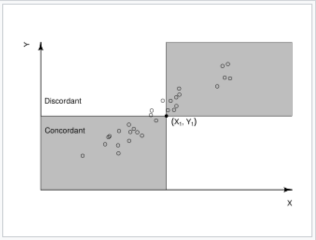


Example 4 – Kendall rank correlation coefficient

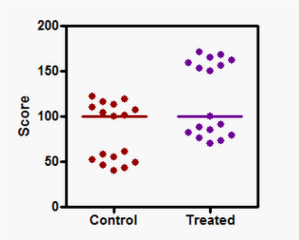
Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of observations of the joint random variables X and Y, such that all the values of (x_i) and (y_i) are unique (ties are neglected for simplicity). Any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be concordant if the sort order of (x_i, x_j) and (y_i, y_j) agrees: that is, if either both $x_i > x_j$ and $y_i > y_j$ holds or both $x_i < x_j$ and $y_i < y_j$; otherwise they are said to be discordant. The Kendall τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})} =$$

$$= 1 - \frac{2\,(\text{number of discordant pairs})}{\binom{n}{2}}$$



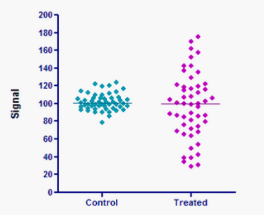
Suppose we have a sample of n_x observations $\{x_1, x_2, \ldots x_n\}$ in one group (i.e. from one population) and a sample of n_y observations $\{y_1, y_2, \ldots y_n\}$ in another group (i.e. from another population).
The Mann-Whitney test is based on a comparison of every observation x_i in the first sample with every observation y_j in the other sample. The total number of pairwise comparisons that can be made is $n_x n_y$.
You'll sometimes read that the Mann-Whitney test compares the medians of two groups. But this is not exactly true, as this example demonstrates.



The graph shows each value obtained from control and treated subjects. The two-tail P value from the Mann-Whitney test is 0.0288, so you conclude that there is a statistically significant difference between the groups. But the two medians, shown by the horizontal lines, are identical. The Mann-Whitney test ranked all the values from low to high, and then compared the mean ranks. The mean of the ranks of the control values is much lower than the mean of the ranks of the treated values, so the P value is small, even though the medians of the two groups are identical.

Mann-Whitney U test

It is also not entirely correct to say that the Mann-Whitney test asks whether the two groups come from populations with different distributions. The two groups in the graph below clearly come from different distributions, but the P value from the Mann-Whitney test is high (0.46). The standard deviation of the two groups is obviously very different. But since the Mann-Whitney test analyzes only the ranks, it does not see a substantial difference between the groups.



The Mann-Whitney test **compares the mean ranks** – it does not compare medians and does not compare distributions. More generally, the P value answers this question: What is the chance that a randomly selected value from the population with the larger mean rank is greater than a randomly selected value from the other population?
If you make an additional assumption – that the distributions of the two populations have the **same shape**, even if they are shifted (have different medians) – then the Mann-Whiteny test can be considered a test of medians. If you accept the assumption of identically shaped distributions, then a small P value from a Mann-Whitney test leads you to conclude that the difference between medians is statistically significant. However, if the groups have the same distribution, then a shift in location will move medians and means by the same amount and so the difference in medians is the same as the difference in means. Thus the Mann-Whitney test is also a test for the difference in means.

Mann-Whitney U test

Hence, when we apply Mann-Whitney test under assumption that the two populations have the **same shape**, we can set our hypotheses as following:

$$H_0 : f_x(x) = f_y(y)$$

$$H_1 : f_x(x) = f_y(y + a)$$

We count the number of times an x_i from sample 1 is greater than a y_j from sample 2 . This number is denoted by U_x. Similarly, the number of times an x_i from sample 1 is smaller than a y_j from sample 2 is denoted by U_y. Under the null hypothesis we would expect U_x and U_y to be approximately equal.
Procedure for carrying out the test:

1. Arrange all the observations in order of magnitude.
2. Under each observation, write down X or Y (or some other relevant symbol) to indicate which sample they are from.
3. Under each x write down the number of y s which are to the left of it (i.e. smaller than it); this indicates $x_i > y_j$. Under each y write down the number of x s which are to the left of it (i.e. smaller than it); this indicates $y_j > x_i$
4. Add up the total number of times $x_i > y_j$ - denote by U_x. Add up the total number of times $y_j > x_i$ - denote by U_y. Check that $U_x + U_y = n_x n_y$.
5. Calculate $U = \min(U_x, U_y)$
6. Use statistical tables for the Mann-Whitney U test to find the probability of observing a value of U or lower. If the test is one-sided, this is your p-value; if the test is a two-sided test, double this probabililty to obtain the p-value.

NOTE: If the number of observations is such that $n_x n_y$ is large enough (¿20), a normal approximation can be used with $\mu_U = \frac{n_x n_y}{2}, \sigma_U = \sqrt{\frac{n_x n_y (N+1)}{12}}$, where $N = n_x + n_y$.

Mann-Whitney U test

Dealing with ties: It is possible that two or more observations nay be the same. If this is the case we can still calculate U by allocating half the tie to the X value and half the tie to the Y value. However, if this is the case then the normal approximation must be used with an adjustment to the standard deviation. This becomes:

$$\sigma_U = \sqrt{\frac{n_x n_y}{N(N-1)} \times \left[\frac{N^3 - N}{12} - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{12}\right]}$$

where

$$N = n_x + n_y$$
$$g = \text{ the number of groups of ties}$$
$$t_j = \text{ the number of tied ranks in group j}$$

Mann-Whitney U test

Example: The following data shows the age at diagnosis of type II diabetes in young adults. Is the age at diagnosis different for males and females?
Males: 19 22 16 29 24; Females: 20 11 17 12
Solution:

1. 1. Arrange in order of magnitude

Age	11	12	16	17	19	20	22	24	29
M/F	F	F	M	F	M	F	M	M	M
$M > F$			2		3		4	4	4
$F > M$	0	0		1		2			

2. 2. Affix M or F to each observation (see above).
3. 3. Under each M write the number of F s to the left of it; under each F write the number of Ms to the left of it (see above).
4. 4. $U_M = 2+3+4+4+4 = 17 \quad U_F = 0+0+1+2 = 3$
5. 5. $U = \min(U_M, U_F) = 3$
6. 6. Using tables for the Mann-Whitney U test we get a two-sided p-value of $p = 0.11$
7. 7. If we use a normal approximation we get:

$$z = \frac{U - \frac{n_x n_y}{2}}{\sqrt{\frac{n_x n_y (N+1)}{12}}} = \frac{3-10}{\sqrt{50/3}} = -1.715 \text{ This gives a two-sided p-value of } p = 0.09.$$

The exact test and the normal approximation give similar results. We would conclude that there is no real evidence that the age at diagnosis is different for males and females, although the results are borderline and the lack of statistical significance in this case may just be due to the very small sample. The actual median age at diagnosis is 14.5 years for females and 22 for males, which is quite a substantial difference. In this case it would be advisable to conduct a larger study.

Welch vs Mann-Whitney

Welch T-test	Mann-Whitney U-test
+ easy to interpret	- difficult to interpret
- sensitive to outliers	+ not sensitive to outliers

It would seem prudent to use non-parametric tests in all cases, which would save one the bother of testing for Normality. Parametric tests are preferred, however, for the following reasons:

1. We are rarely interested in a significance test alone; we would like to say something about the population from which the samples came, and this is best done with estimates of parameters and confidence intervals.
2. It is difficult to do flexible modelling with non-parametric tests, for example allowing for confounding factors using multiple regression.
3. Parametric tests usually have more statistical power than their non-parametric equivalents. In other words, one is more likely to detect significant differences when they truly exist.