Difference

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

Before Inferential Statistics

- We need to establish evidence to support going ahead with building a predictive model.
- If we are asserting a relationship,
  - we need to investigate if there is any evidence of a relationship using the appropriate test and make a decision based on the results (strength, direction, etc.).
- If we are asserting a differential effect for different groups,
  - we need to investigate if there is any difference using the appropriate test and make a decision based on the result.

Relationships and Causality

- How to determine if a relationship is causal?
  - Identify the directionality of relationships between **independent** (also called predictor variable) and **dependent variables** (also called outcome variable) based on research/prior knowledge.
- **Independent Variable** hypothesised to cause changes in **Dependent Variable**.

Examples

Independent Variable	**Dependent Variable**
Death Penalty	Crime Rate
School Funding	Graduation Rate
Graduation Rate	Crime Rate
Parental Income	Graduation Rate
Parental Education Level	Graduation Rate
Revision Time	Exam Result

Table: Some examples for independent and dependent variables in different experiments.

Causal Relationship

- To assert a causal relationship is to claim that changes in the independent variable create changes in the dependent variable.
- In practice can only assert that one factor causes change in another when you can satisfy the following criteria:
  - Association: There must be a relationship between the two variables.
  - Time Order: The change in the independent variable *precedes* the change in the dependent.
  - Non-spuriousness: The effects of a third unmeasured 'spurious' factor does not produce the relationship between the two variables.

Causal Relationships - Association

- Variables correspond to each other in predictable ways.
    - Can be either positive (an increase in one causes an increase in the other),
    - Or negative (an increase in one causes a decrease in the other).
- Many associations are not causal, e.g.:
    - Red cars do not cause more accidents.
    - Rather aggressive drivers are more likely to buy red cars.
    - Aggressive drivers are more likely to be involved in accidents.

Causal Relationship - Time Order

- Logically when a change in one variable causes a change in another, the change in the independent needs to occur before the change in the dependent.
- Need to be careful ...
    - Even if time order is satisfied, a causal relationship may not exist.

Causal Relationship - Non-spuriousness

- A spurious relationship exists when two variables appear to be causally related but their relationship can be attributed to a third unidentified variable.
- E.g.:
  - Ice cream consumption causing drowning deaths.
  - There is an association between ice cream consumption and drowning (more people die during times when a lot of ice cream is being consumed).
  - Time order can also be satisfied - increase in ice cream sales precede increase in drowning deaths.
  - However, the unmeasured factor here is seasonal temperature:
    - In warm weather more people consume ice cream and swim.

Relationships and Causality

- Causality:
  - Determining how particular sets of conditions (represented by variables) lead to particular outcomes (also represented by variables).
- Seldom are relationships **deterministic**:
  - Everyone completing a university degree will have a higher income than those who do not (unlikely to be true).
- Most are **probabilistic**:
  - i.e. factors increase or decrease tendency towards particular outcomes.
  - Those with a university degree tend to earn more than those who do not.
  - University education is a factor that pushes another factor (income) in a predictable direction.

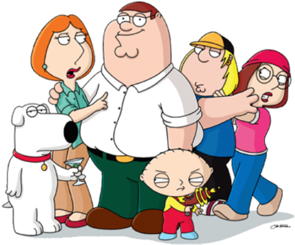
Difference

Comparison between Samples

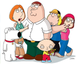
Sometimes we want to investigate if an effect is different for different groups within our population.

There are a range of statistical tests we can use. The choice depends on:

- The measurement of the variable.
- The shape of the data.



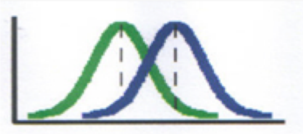
Tests for Group Comparison

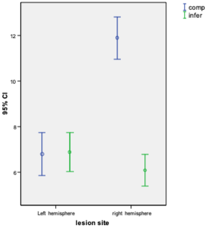
	Independent Samples	**Related Samples** also called dependent means test
Interval measures/ parametric	Independent samples t-test* 	Paired samples t-test**
Ordinal/non-parametric	Mann-Whitney U-Test	Wilcoxon test

* 2 different groups of participants

** 2 same participants measured at two different points

Parametric: t-tests





- Compare the **mean** between 2 samples/groups/conditions.
- **If 2 samples are taken from the same population, then they should have fairly similar means → if 2 means are statistically different, then the samples are likely to be drawn from 2 different populations**, i.e. they really are different

t Distribution

- Very similar to the Z distribution by assuming normality.
- Normality is obtained after about 120 data observations.
    - In which case $t = z$.
- Basic rule of parameter estimation:
    - The higher the observations (N) of sample, the more reflective of overall population.

The t Distribution

- The t distribution is a short, fat relative of the normal.
- The shape of t depends on its degrees of freedom.
- As N becomes infinitely large, t becomes normal and becomes z.

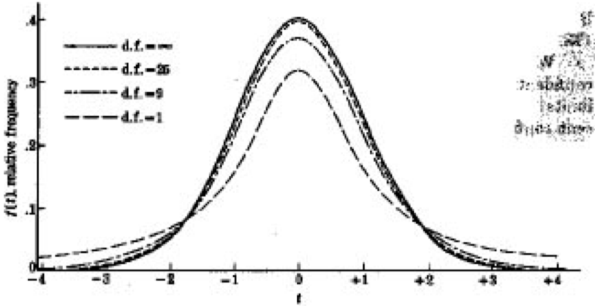


fig. 10.1 Distribution of *t* for various degrees of freedom. (*From D. Lewis, quantitative methods in psychology, McGraw-Hill Book Company, New York, 1960.*)

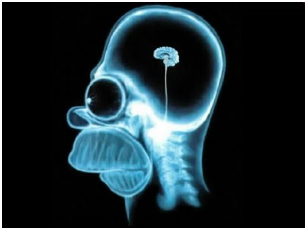
t distribution critical values table

- Only need to check if your sample is smaller than 120.
- `https://www.easycalculation.com/statistics/t-distribution-critical-value-table.php`
- Your statistical tool will do this look up for you and calculate whether your test is statistically significant or not.

Comparison between Conditions

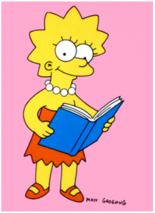
Is the activation (from MRI) different when you compare 2 different conditions?

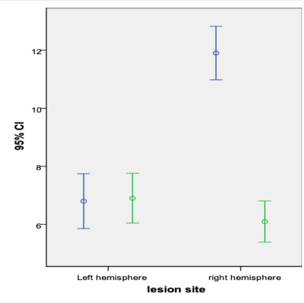
Reading aloud (script) vs "Reading" finger spelling (sign)



Reading aloud vs Picture naming

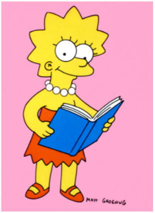
t-test

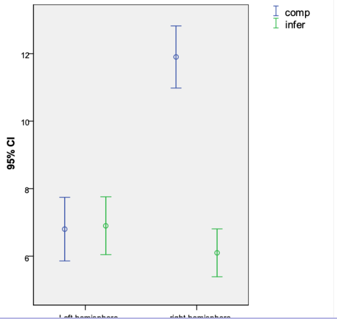




- **Exp. 1**
- Reading script (blue) is compared to ”reading” finger spelling (green).
- Activation patterns are similar, not significantly different → they are similar tasks.

t-test





- **Exp. 2**
- When picture naming (green) is compared to reading aloud (blue) those exact object labels (e.g. naming the picture of a tiger versus reading the word ”tiger”).
  - Reading causes significantly stronger activation and this is different to naming.
  - Activation patterns are very (and significantly) different → reading aloud

Formula for test statistic

Difference between the means divided by the pooled standard error of the mean.

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_{\overline{x_1} - \overline{x_2}}} \tag{1}$$

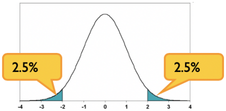
SE variability between sample means

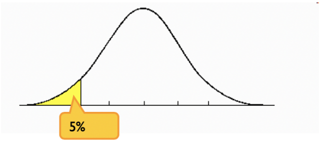
If large then may be a problem with representativeness of the sample.

Reporting convention: $t = 11.456, df = 9, p < 0.001$ (Remember *df* depends on the number in your sample).

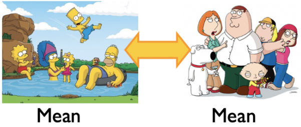
Types of t-tests cont.

2-tailed tests vs one-tailed tests



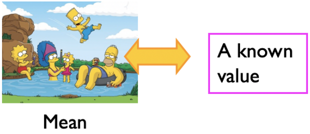


2 sample t-tests 1 sample t-tests



Mean

Mean



Mean

A known value

Assumptions

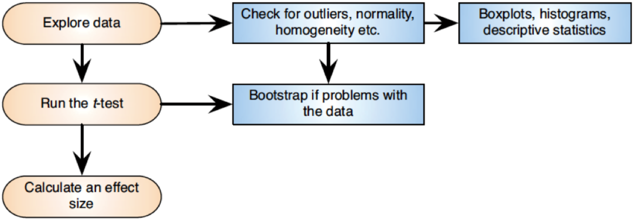
- The t-test is based on assumptions of normality and homogeneity of variance.
    - You can test for both of these.
- As long as the samples in each group are large and nearly equal, the t-test is robust, that is, still good, even though assumptions are not met.

Assumptions of the t-test

- Both the independent t-test and the dependent t-test are parametric tests based on the normal distribution.
- Therefore, they assume:
    - Data are measured at least at the interval level.
    - The sampling distribution is normally distributed.
    - In the dependent t-test this means that the sampling distribution of the differences between scores should be normal, not the scores themselves.
- The independent t-test, because it is used to test different groups of people, also assumes:
    - Variances in these populations are roughly equal (homogeneity of variance).
    - Scores in different treatment conditions are independent (because they come from different people).

Independent t-test using R

FIGURE 9.3 The general process for performing a *t*-test



Explore data

Check for outliers, normality, homogeneity etc.

Boxplots, histograms, descriptive statistics

Run the *t*-test

Bootstrap if problems with the data

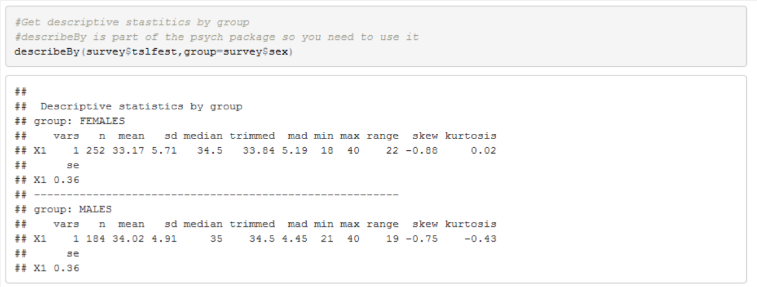
Calculate an effect size

Independent t-test Example

- Survey.dat (Julie Pallant)
- Dataset created from a survey designed to explore the factors that impact on respondents' psychological adjustment and wellbeing.
- Question:
  - Is there a significant difference in the mean self-esteem scores for males and females?
- Need:
  - One categorical variable (male/female gender).
  - One continuous, dependent variable (self-esteem score tslfest).
- T-test:
  - Will tell you whether there is a statistically significant difference between the mean scores of the groups.

Independent t-test in R

We need to look at the descriptive statistics for the variable by group.

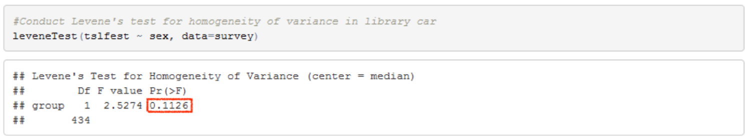


```
#Get descriptive stastitics by group
#describeBy is part of the psych package so you need to use it
describeBy(survey$tslfest,group=survey$sex)
```

```
##
##  Descriptive statistics by group
## group: FEMALES
##    vars   n  mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 252 33.17 5.71   34.5   33.84 5.19  18  40    22 -0.88     0.02
##      se
## X1 0.36
## --------------------------------------------------------
## group: MALES
##    vars   n  mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 184 34.02 4.91     35    34.5 4.45  21  40    19 -0.75    -0.43
##      se
## X1 0.36
```

Independent t-test in R

We need to know the homogeneity of variance in advance so we can set the parameter on the t-test so we conduct the Levene's test first.



```
#Conduct Levene's test for homogeneity of variance in library car
leveneTest(tslfest ~ sex, data=survey)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  2.5274 0.1126
##       434
```

The null hypothesis is that all variances are equal. A resulting p-value under 0.05 means that variances are not equal and than further parametric tests are not suitable. Note that this test is meant to be used with normally distributed data but can tolerate relatively low deviation from normality.

Testing for Homogeneity of Variance

In R you need to choose and test in advance of running other tests:

- Levene's test is best - this is the most robust test for normally distributed data. Null hypothesis is that the variance is equal.
- Fligner-Killeen test - this is a non-parametric test equivalent.
- To run in R (Levene's is in the library car):

```r
library(car)
leveneTest(continuousvariable ~ categoricalvariable,
    data=dataframename)
```

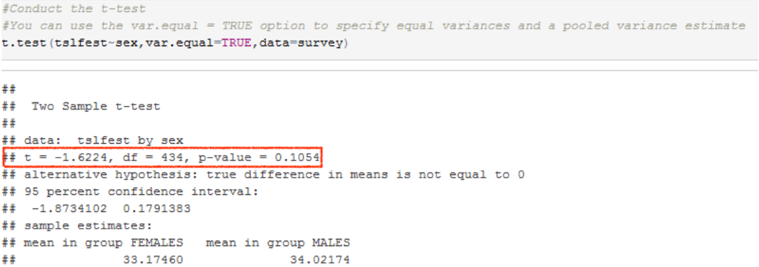
- E.g. $leveneTest(tslfest \sim sex, data = sdata)$
- Gives as output an F statistic and an estimate of significance

Independent t-test in R

Conduct the t-test: **t.test()** function produces a variety of t-tests.

```
t.test(y~x) # where y is numeric and x is a binary
    factor
```

- Unlike most statistical packages, the default assumes unequal variance and applies the Welsh df modification.
- You can use the *var.equal* = *TRUE* option to specify equal variances.

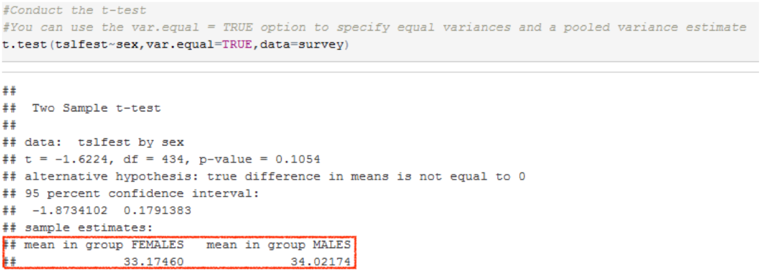


```
#Conduct the t-test
#You can use the var.equal = TRUE option to specify equal variances and a pooled variance estimate
t.test(tslfest~sex,var.equal=TRUE,data=survey)
```

```
##
##  Two Sample t-test
##
## data:  tslfest by sex
## t = -1.6224, df = 434, p-value = 0.1054
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.8734102  0.1791383
## sample estimates:
## mean in group FEMALES   mean in group MALES
##              33.17460              34.02174
```

Independent t-test in R

Look also at the mean difference:



```
#Conduct the t-test
#You can use the var.equal = TRUE option to specify equal variances and a pooled variance estimate
t.test(tslfest~sex,var.equal=TRUE,data=survey)
```

```
##
##  Two Sample t-test
##
## data:  tslfest by sex
## t = -1.6224, df = 434, p-value = 0.1054
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.8734102  0.1791383
## sample estimates:
## mean in group FEMALES   mean in group MALES
##              33.17460              34.02174
```

Calculating effect size

- Cohen's d
  - Value for a between subjects t-test and the degrees of freedom.

$$d = \frac{2t}{\sqrt{df}} \tag{2}$$

In this case $d = 0.155$.

Reporting Cohen's d:

$0.2 = small\ effect, 0.5 = moderate, 0.8 = large$

NOTE

Choice of using eta squared or Cohen's d depends on your field of study.

Calculating effect size

- Magnitude of the difference between the means of your groups.
- Eta squared ranges from 0 to 1.

$$\textit{eta squared} = \frac{t^2}{t^2 + (N_1 + N_2 - 2)} \tag{3}$$

```
    T = 1.62
    N1 = Number in group 1 (males) 184
    N2 = number in group 2 (females) 252

```

$$\textit{eta squared} = \frac{1.62^2}{1.62^2 + (184 + 252 - 2) = 0.006} \tag{4}$$

```
    Guidelines on effect size: 0.01 = small, 0.06 = moderate, 0.14 = large

```

Reporting a t-test

"An independent-samples t-test was conducted to compare self-esteem scores for male and female respondents. No significant difference in the scores for self-esteem was found ($M = 33.17, SD = 5.71$ for female respondents, $M = 34.02, SD = 4.91$ for male respondents), ($t(434) = -1.62, p = 0.10$). The eta square statistic also indicated a very small effect size (0.006)."

Assumptions for Non-Parametric Tests

- Random samples
- Independent observations
    - Each case is counted only once.
    - Unless it is a repeated measure where the same participants are retested on different occasions or under different conditions.

Non-Parametric Example from Andy Field

- Our neurologist investigating the depressant effects of certain alcoholic drinks:
  - Tested 20 clubbers.
  - 10 were allowed to drink only vodka on a Saturday night.
  - 10 were allowed to drink only beer.
  - Levels of depression were measured using the Beck Depression Inventory (BDI) the day after and midweek.
- We are hypothesising that two groups of different entities will differ on the result of a test for depression.
- A similar number of high and low ranks in each group would suggest depression levels do not differ between the groups.
- A greater number of high ranks in the vodka group than the beer group would suggest the vodka group is more depressed than the beer group.
- **Field-BDI-NonParametric.dat**

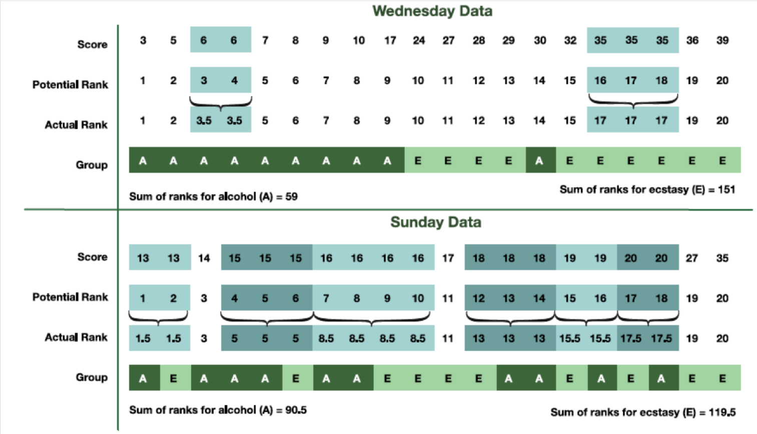
Non-Parametric: Comparing two independent conditions

- This requires us to fit a model that compares the distribution of this variable for those in the beer group and the vodka group.
- Requires us to test differences between two conditions in which different participants have been used.
- To compare the distribution of two conditions and these differences, you can use:
  - The Mann-Whitney test and
  - Wilcoxon's Rank-sum test.
- These work on ranking data.

Ranking Data

- The tests work on the principle of ranking the data for each group:
  - Lowest score = a rank of 1,
  - Next highest score = a rank of 2, and so on.
  - Tied ranks are given the same rank: the average of the potential ranks.
- For an unequal group size the test statistic (Ws) = sum of ranks in the group that contains the least people.
- For an equal group size the test statistic Ws = the value of the smaller summed rank.
- Add up the ranks for the two groups and take the lowest of these sums to be our test statistic.
- The analysis is carried out on the ranks rather than the actual data.

Ranking the Depression scores for Wednesday and Sunday



Wednesday Data

Score	3	5	6	6	7	8	9	10	17	24	27	28	29	30	32	35	35	35	36	39
Potential Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Actual Rank	1	2	3.5	3.5	5	6	7	8	9	10	11	12	13	14	15	17	17	17	19	20
Group	A	A	A	A	A	A	A	A	A	E	E	E	E	A	E	E	E	E	E	E

Sum of ranks for alcohol (A) = 59

Sum of ranks for ecstasy (E) = 151

Sunday Data

Score	13	13	14	15	15	15	16	16	16	16	17	18	18	18	19	19	20	20	27	35
Potential Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Actual Rank	1.5	1.5	3	5	5	5	8.5	8.5	8.5	8.5	11	13	13	13	15.5	15.5	17.5	17.5	19	20
Group	A	E	A	A	A	E	A	A	E	E	E	E	A	A	E	A	E	A	E	E

Sum of ranks for alcohol (A) = 90.5

Sum of ranks for ecstasy (E) = 119.5

FIGURE 6.4 Ranking the depression scores

Provisional analysis using R

- Load the data:
  $drinkset < -read.table('' Field - BDI - Non - parametric.dat'')$
- Get your descriptive statistics:
  $describeBy(drinkset\$bdisun, group = drinkset\$drink)$
- Create your plots using ggplot

Mann-Whitney test

- It is used to test the null hypothesis that two samples come from the same population (i.e. have the same median)
- OR alternatively, whether observations in one sample tend to be larger than observations in the other.
- Compares the medians from two populations and works when the Y variable is continuous and the X variable is discrete with two attributes.
- Of course, the Mann-Whitney test can also be used for normally distributed data, but in that case it is less powerful than the 2-sample t-test.

The U Value

- The U value for each group is calculated by subtracting the possible minimum rank which the group can take from the sum of the ranks, and the smallest U value is used for the test.
- For the test of significance of the Mann-Whitney U-test it is assumed that with $n > 80$ or each of the two samples at least > 30 the distribution of the U-value from the sample approximates normal distribution.
- U value is compared to table of critical values depending on the size of each sample
- U value must be equal to or less than critical value.

In R - Mann Whitney U (Wilcoxon Test)

```
# From stats package
wilcox.test(y~A)
# where y is your continuous variable and A is your categorical
drinkset<-read.table(’Field-BDI-Non-parametric.dat’)
wilcox.test(bdisun~drink, data=drinkset)

Wilcoxon rank sum test with continuity correction data: bdisun by drink
W = 35.5, p-value = 0.2861 alternative hypothesis: true location shift is not equal to 0

wilcox.test(bdiwed~drink, data=drinkset)

Wilcoxon rank sum test with continuity correction data: bdiwed by drink
W = 4, p-value = 0.000569 alternative hypothesis: true
```

Calculating an Effect Size

The equation to convert a z-score into the effect size estimate, r, is as follows (from Rosenthal, 1991: 19)[1]:

$$r = \frac{Z}{\sqrt{N}} \tag{5}$$

- z is something we need to calculate if we use the stats package.
- We can use *qnorm*() of the *pvalue* to get it.

[1] Rosenthal 1991 Meta-analytic procedures for social research, SAGE

In R - Calculate effect size

Calculate the standardised z statistic Z and call it Zstat

$Zstat < -qnorm(test\$p.value/2)$

Calculate the effect size using $abs(Zstat)/sqrt(20)$

$\#$ 20 is the number in the study
$abs(Zstat)/sqrt(20)$

$$1$$

0.8737202

In R - Mann Whitney U

From coin package:

```
wilcox.test(y~A)
# where y is your continuous variable and A is your categorical
drinkset<-read.table(’Field-BDI-Non-parametric.dat’)
Coin::Wilcox_test(bdisun~drink, data=drinkset)

Asymptotic Wilcoxon-Mann-Whitney Test data: bdisun by drink (Beer, Vodka)
Z = -1.1049, p-value = 0.2692
alternative hypothesis: true mu is not equal to 0

Wilcox_test(bdiwed~drink, data=drinkset)
Asymptotic Wilcoxon-Mann-Whitney Test data: bdiwed by drink (Beer, Vodka)
Z = -3.4838, p-value = 0.0004943
alternative hypothesis: true mu is not equal to 0
```

Calculating an Effect Size

The equation to convert a z-score into the effect size estimate[2], r:

$$r = \frac{Z}{\sqrt{N}} \tag{6}$$

- We need to calculate z if we use the stats package.
- N is the size of the study (number of total observations).
- 10 vodka and 10 beer users = 20 total observations.

$$r_{Sunday} = \frac{-1.11}{\sqrt{20}} = -0.25 \tag{7}$$

$$r_{Wednesday} = \frac{-3.48}{\sqrt{20}} = -0.78 \tag{8}$$

Rounded value of Z.

[2] Rosenthal 1991 Meta-analytic procedures for social research, SAGE

Reporting the Results

For the Mann-Whitney test

Depression levels in vodka users ($Mdn = 17.50$) did not differ significantly from beer users ($Mdn = 16.00$) on the day after the drugs were taken, ($U = 35.50, z = -1.11, p = .280, r = -.25$). However, by Wednesday, vodka users ($Mdn = 33.50$) were significantly more depressed than beer users ($Mdn = 7.50$), ($U = 4.00, z = -3.48, p < .001, r = -.78$).

NOTE

Our W statistic is what we use for U. We report the median values because the data is not normally distributed.

Repeated Measures

Parametric: Paired samples t-test

- Also referred to as **repeated measures** test.
- Collect data from one group on two different occasions or under different conditions.

Paired Samples t-test

- Using Pallant's experim.dat.
- Investigating the impact of an intervention designed to increase student confidence in their ability to survive a statistics course.
- Students completed a Fear of Statistics Test at Time 1 (FOST1) (before the intervention) and Time 2 (FOST2) (after the intervention).
- Question:
  - Is there a significant change in participant's Fear of Statistics Test scores following participation in an intervention designed to increase students confidence in their ability to complete a statistics course?
  - Does this intervention have an impact on participants' Fear of Statistics Test scores?

Paired Samples t-test

- Need:
  - One categorical independent variable (in this case Time with two different levels Time 1 and Time 2).
  - One continuous dependent variable (Fear of Statistics score) measured on two different occasions or under different conditions (FOST1 and FOST2).
- What will a paired samples t-test do?
  - Will tell you whether there is a statistically significant difference between the mean scores for Time 1 and Time 2.
- Non-parametric alternative = Wilcoxon Signed Rank Test.

In R

Paired T-test:

```
edata<-read.table(’experim.dat’)

```

For our example:

```
t.test(edata$fost1,edata$fost2,paired=TRUE)

```

Output:

```
data: edata$fost1 and edata$fost2 t = 5.3936, df = 29, p-value = 8.498e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 1.655478 3.677855
sample estimates: mean of the differences 2.666667

```

Interpreting our output

```
t = 5.3936, df = 29, p-value = 8.498e-06
sample estimates: mean of the differences 2.666667
```

- We have established a significant difference to the level of 2.67 (rounded).
- Having established a significant difference, establish which set of scores are higher.
- Look at Paired Sample Statistics.
    - Mean @ Time 1 $=$ 40.17, Mean @ Time 2 $=$ 37.50.
    - Conclude there was a significant decrease from Time 1 to Time 2.
    - We cannot however conclude that the intervention caused this!

Calculating an Effect Size

$$\textit{eta squared} = \frac{t^2}{t^2 + (N-1)} = \frac{(5.39)^2}{(5.39)^2 + (30-1)} = 0.5 \tag{9}$$

Cohen's guidelines:

- .01 = small
- .06 = moderate
- .14 = large

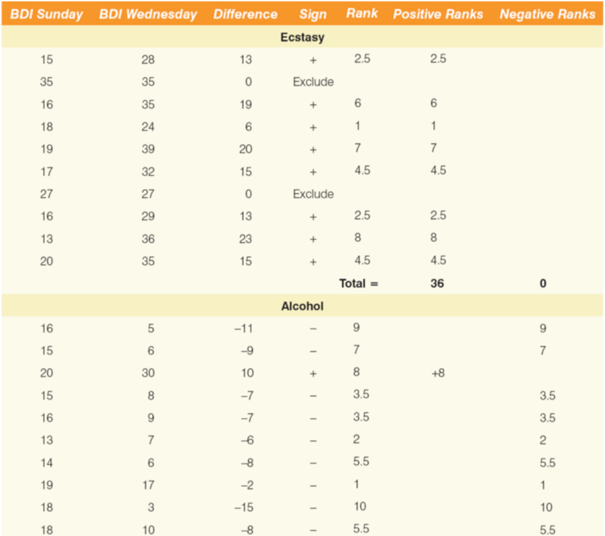
Reporting paired t-test

"A paired-samples t-test was conducted to evaluate the impact of the intervention on students' scores on the Fear of Statistics Test (FOST). There was a statistically significant decrease in FOST scores from Time 1 ($M = 40.17, SD = 5.16$) to Time 2 ($M = 37.5, SD = 5.15, t(29) = 5.29, p < .001$). The mean decrease in FOST scores was 2.67 with a 95% confidence interval ranging from 1.66 to 3.68. The eta squared statistic (.50) indicated a large effect size."

Non-Parametric Data: the Wilcoxon signed-rank test

- Uses:
  - To compare two sets of scores, when these scores come from the same participants.
- Using our data on clubbers FieldBDINonParametric.dat.
- Imagine the experimenter was interested in the change in depression levels for each drink.
  - Non-parametric test because the distributions of scores for both drinks were non-normal on one of the days.
- Only want the scores for vodka.
  - Need to select the cases for vodka.
  - Or split the file.

Ranking data in the Wilcoxon signed-rank test



BDI Sunday	BDI Wednesday	Difference	Sign	Rank	Positive Ranks	Negative Ranks
			Ecstasy			
15	28	13	+	2.5	2.5	
35	35	0	Exclude			
16	35	19	+	6	6	
18	24	6	+	1	1	
19	39	20	+	7	7	
17	32	15	+	4.5	4.5	
27	27	0	Exclude			
16	29	13	+	2.5	2.5	
13	36	23	+	8	8	
20	35	15	+	4.5	4.5	
				Total =	**36**	**0**
			Alcohol			
16	5	−11	−	9		9
15	6	−9	−	7		7
20	30	10	+	8	+8	
15	8	−7	−	3.5		3.5
16	9	−7	−	3.5		3.5
13	7	−6	−	2		2
14	6	−8	−	5.5		5.5
19	17	−2	−	1		1
18	3	−15	−	10		10
18	10	−8	−	5.5		5.5

In R dependent 2-group Wilcoxon Signed Rank Test

Need to split the file or subset it and run against each sub-set:

```
vodkadata<-subset(drinkset,drinkset$drink=='Vodka')
beerdata<-subset(drinkset,drinkset$drink=='Beer')
```

Wilcoxon test syntax:

```
wilcox.test(y1,y2,paired=TRUE)
```

Example:

```
stats::wilcox.test(beerdata$bdisun,beerdata$bdiwed,
    paired=TRUE)
```

Output

```
Wilcoxon signed rank test with continuity correction
data: beerdata$bdisun and beerdata$bdiwed
V = 47, p-value = 0.05248 alternative hypothesis: true
    location shift is not equal to 0
```

The value $V = 0$ corresponds to the sum of ranks assigned to the

Calculating an Effect size

- The effect size can be calculated in the same way as for the Mann-Whitney test.
- Again calculate Z:
  - For the vodka group z is 2.53, and for the beer group is -1.99.
- In both cases we had 20 observations (although we only used 10 people and tested them twice, it is the number of observations, not the number of people, that is important here).
- The effect size is therefore $\frac{2.53}{\sqrt{20}} = 0.57$, $\frac{-1.99}{\sqrt{20}} = -0.44$.

Reporting the results

Reporting Test-Statistic

For vodka users depression levels were significantly higher on Wednesday ($Mdn = 33.50$) than on Sunday ($Mdn = 17.50$), $T = 36, p = .012, r = 0.57$. However, for beer users the opposite was true: depression levels were significantly lower on Wednesday ($Mdn = 7.50$) than on Sunday ($Mdn = 16.0$), $T = 8, p = .047, r = -0.44$.

Reporting the results

Reporting the values of z

For vodka users, depression levels were significantly higher on Wednesday ($Mdn = 33.50$) than on Sunday ($Mdn = 17.50$), $z = 2.53, p = .012, r = 0.57$. However, for beer users the opposite was true: depression levels were significantly lower on Wednesday ($Mdn = 7.50$) than on Sunday ($Mdn = 16.0$), $z = -1.99, p = .047, r = -0.44$.

Back to Hypothesis Testing

Errors in Hypothesis Tests

- Just because we find a statistically significant difference/effect does not necessarily indicate there is a causal relationship.
- Because the hypothesis test relies on sample data, and because sample data are not completely reliable, there is always the risk that misleading data will cause the hypothesis test to reach a wrong conclusion.

Type I Errors

- Occur when the sample data appear to show an effect/difference when, in fact, there is none in the population.
- In this case the researcher will reject the null hypothesis and falsely conclude that there is an effect/difference.
- Type I errors are caused by unusual, unrepresentative samples.
- Just by chance the researcher selects an extreme sample with the result that the sample falls in the critical region even though there is no effect.
- The hypothesis test is structured so that Type I errors are very unlikely.
- Specifically, the probability of a Type I error is equal to the alpha level.

Type II Errors

- Occurs when the sample does not appear to have an effect/difference when in fact this exists in the population.
- In this case, the researcher will fail to reject the null hypothesis.
- Type II errors are commonly the result of a very small effects/differences (not large enough to show up in the research study).

Power of a Hypothesis Test

- The **power** β of a hypothesis test is defined is the probability that the test will reject the null hypothesis when there is no effect.
- The power of a test depends on a variety of factors including the size of the effect and the size of the sample.

Possible Error. . . ?

- Compare Type I and Type II error like this:
  - The only concern when you find statistical significance ($p < 0.05$) is Type I Error.
    - Is the difference between groups REAL or due to Random Sampling Error.
    - Thankfully, the p-value tells you exactly what the probability of that random sampling error is.
    - In other words, the p-value tells you how likely Type I error is.
- But, does the p-value tell you how likely Type II error is?
  - The probability of Type II error is better provided by Power.

Possible Error. . . ?

- Probability of Type II error is provided by Power:
  - Statistical Power, also known as β (Beta) (actually $1 - \beta$)
- Power (Beta) is related to Alpha, but:
  - Alpha is the probability of having Type I error.
  - Lower number is better (i.e., 0.05 vs 0.01 vs 0.001).
  - Power is the probability of NOT having Type II error.
  - The probability of being right (correctly rejecting the null hypothesis).
  - Higher number is better (typical goal is 0.8).

Should it be statistically significant?

The most obvious thing you need to consider is if you REALLY should have found a statistically significant result?

- Just because you wanted your test to be significant doesn't mean it should be.
- This wouldn't be Type II error – it would just be the correct decision.

15 min break

Lab Exercise

- Load the Regression.sav dataset.
- Describe the variable *normexam* by group *girl* (= gender).
- Check for homogeneity of variance.
- Check for difference in the two groups.
- Optional: Check for gender differences in the *survey.dat* dataset.