

Comparing More Than Two Groups and Non-numerical Variables

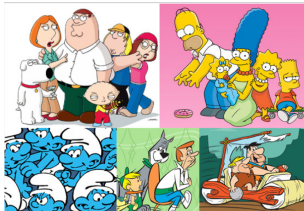
Probability and Statistical Inference

Bojan Božić

TU Dublin



Winter 2020

Comparison of more than 2 samples



To find the difference between these groups, we will use ANOVA.

Tests for Multiple Group Comparison

	Independent Samples	Related Samples
Interval measures/ parametric	ANOVA ¹ 	Repeated ² Measures ANOVA 
Ordinal/non-parametric	Kruskal-Wallis	Friedman

¹multiple different groups of participants

²multiple same participants measured at multiple different points

ANOVA

- **AN**alysis **Of** **VA**riance (**ANOVA**)
 - Still compares the differences in means between groups but it uses the variance of data to "decide" if means are different.
 - Really is **ANOVASMD** (**A**nalysis **of** **v**ariance to **s**ee if **m**ean are **d**ifferent).
- Looks to see what the variation (variance) is *within* the groups, then works out how that variation would translate into variation (i.e. differences) *between* the groups, taking into account how many subjects there are in the groups.
- If the observed differences are a lot bigger than what you'd expect by chance, you have statistical significance.

ANOVA

- Factors: the overall 'things' being compared (e.g. age, task, score).
- Levels: the elements of the factor (young vs old and naming vs reading aloud).
- F-statistic or F-ratio:
 - Magnitude of the difference between the different conditions.
 - Similar to z or t-score as it compares the amount of systematic variance in the data to the amount of unsystematic variance.
 - It is the ratio of the experimental effect to the individual differences in performance.
 - Less than 1, it must represent a non-significant event (so you always want a F-ratio greater than 1)
 - Degrees of freedom – depends on the number of factors and the number of levels

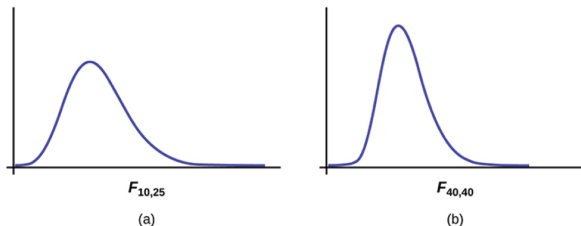
ANOVA

- **ANOVA tests for one overall effect only.**
 - **Omnibus test.**
- There is a need for post-hoc testing.
 - ANOVA can tell you if there is an effect but not where.
 - Need to perform multiple comparisons to identify level of effect and where effect is evident.
- To test for significance,
 - obtained F-ratio is compared against maximum value one would expect to get by chance alone in an F-distribution with the same degrees of freedom.
 - p-value associated with F is probability that differences between groups could occur by chance if null-hypothesis is correct.

Reporting convention: $F = 65.58, df = 4, 45, p < .001$

F Distribution

- The curve is not symmetrical but skewed to the right.
- There is a different curve for each set of degrees of freedom (*dfs*).
- The F statistic is greater than or equal to zero.
- As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.



What Does ANOVA Tell us?

- Null Hypothesis:
 - Like a t-test, ANOVA tests the null hypothesis that the means of the different groups are the same.
- Alternate Hypothesis:
 - The means differ.
- ANOVA is an Omnibus test.
 - It test for an overall difference between groups.
 - It tells us that the group means are different.
 - It doesn't tell us exactly which means differ.

One-way Between-Groups ANOVA

- survey.dat (Julie Pallant)
- Question:
 - Is there a difference in optimism scores for young, middle-aged and old participants?
- Need:
 - One independent variable with three or more levels (age category) (agegp3).
 - One continuous variable (optimism scores - toptim).
- Non-parametric equivalent - Kruskal-Wallis Test.

One-way Between-Groups ANOVA

- One-way ANOVA will tell us whether there are significant differences in optimism scores (`toptim`) across the three groups (`agegp3`).
- It will just do that.
 - Tell us if there is a difference but not where the difference is.
- We need to conduct post-hoc tests (additional tests after the ANOVA) to find out where the differences lie.
 - E.g. Is it between young and old, young and middleaged, middleaged and old?

Note

We consider `toptim` as we would any scale variable in advance to see if it can be considered normal, it can.

Preliminaries

Summary statistics for the variable of interest for each group:

```
library( dplyr )  
group_by( sdata , sdata$agegp3 ) %>%  
summarise(  
  count = n() ,  
  mean = mean(toptim , na.rm = TRUE) ,  
  sd = sd(toptim , na.rm = TRUE) )
```

Testing for homogeneity of variance in R

- We need to check for homoscedasticity for the variable of interest across groups.
- In R we use Bartlett's test.

```
bartlett.test(sdata$stoptim, sdata$agegp3)
```

Can be argued that the variances are homogeneous if the $p - value > 0.05$.

Output:

```
Bartlett's K-squared=1.4561,  
df=2, p-value=0.4828
```

Basic ANOVA

```
# Compute the analysis of variance  
res.aov <- aov(sdata$toptim ~ sdata$agegp3,  
data = sdata)  
# Summary of the analysis  
summary(res.aov)
```

```
> summary(res.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sdata\$agegp3	2	179	89.53	4.641	0.0101 *
Residuals	432	8334	19.29		

```
---  
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
4 observations deleted due to missingness
```

There is a statistically significant difference.

But where is the difference?

TukeyHSD(res.aov)

```
> TukeyHSD(res.aov)
Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = sdata$stoptim ~ sdata$agegp3, data = sdata)

`sdata$agegp3`
      diff      lwr      upr      p adj
30 - 44-18 - 29 0.7440309 -0.4489781 1.937040 0.3080109
45+-18 - 29     1.5950113  0.3636472 2.826376 0.0069296
45+-30 - 44     0.8509804 -0.3687705 2.070731 0.2296685
```

We can see the only significant difference is between our 45+ age group and our 18-29 age group and that the magnitude of difference is 1.6.

Testing for homogeneity of variance

What we need to know so we can conduct the correct post-hoc test:

- Homocedasticity – Tukey's honestly significant difference (HSD) post hoc test.
- Otherwise use Games-Howell.
- In R need to specify in relevant post-hoc test function.

In R – One Way ANOVA

```
#read the data into a dataframe  
sdata <-read.table('survey.dat')
```

```
#install and load the library userfriendlyscience  
#it has a really nice one-way anova function that  
#provides nice summary output  
library(userfriendlyscience)
```

```
#run a one-way anova test using the correct  
#post-hoc test Tukey in our case  
one.way <- oneway(sdata$agegp3, y = sdata$toptim,  
posthoc = 'Tukey')
```

```
#printout a summary of the anova
```


In R – One Way ANOVA

Output:

Omega squared: 95% CI = [0; .05], point estimate = .02
Eta Squared: 95% CI = [0; .05], point estimate = .02

	SS	Df	MS	F	p
Between groups (error + effect)	179.07	2	89.53	4.64	.01
Within groups (error only)	8333.95	432	19.29		

Post hoc test: Tukey

	diff	lwr	upr	p adj
30 - 44-18 - 29	0.74	-0.45	1.94	.308
45+-18 - 29	1.6	0.36	2.83	.007
45+-30 - 44	0.85	-0.37	2.07	.230

We can see the only significant difference is between our 45+ age group and our 18-29 age group and that the magnitude of difference is 1.6.

Calculating the effect size

- $\eta^2 = \text{sum of squares between groups} / \text{total sum of squares}$ (from our ANOVA output (rounded up))
- $= 179 / 8334$
- $= .02$

Guidelines on effect size

0.01 = small, 0.06 = moderate, 0.14 = large

Reporting the results

A one-way between-groups analysis of variance was conducted to explore the impact of age on levels of optimism, as measured by the Life orientation Test (LOT). Participants were divided into three groups according to their age (Group 1: 29 yrs or less; Group 2: 30 to 44 yrs; Group 3: 45 yrs and above). There was a statistically significant difference at the $p < .05$ level in LOT scores for the three age groups: $F(2, 432) = 4.6, p = .01$. Despite reaching statistical significance, the actual difference in mean scores between groups was quite small. The effect size, calculated using eta squared was .02. Post-hoc comparisons using the Tukey HSD test indicated that the mean score for Group 1 ($M = 21.36, SD = 4.55$) was statistically different to Group 3 ($M = 22.96, SD = 4.49$). Group 2 ($M = 22.10, SD = 4.15$) did not differ significantly from either Group 1 or 3.

What is this Bonferroni Correction?

- When you conduct multiple analyses between groups on the same dependent variable you are inflating the chance of a Type I error.
 - Incorrectly rejecting the null hypothesis.
- Bonferroni corrects/adjusts the p value by dividing the original α -value by the number of analyses on the dependent variable.

Beware! Multiple Comparison Problem

- When reading a study, you can only account for multiple comparisons when you know about all the comparisons made by the investigators.
- If they report only "significant" differences, without reporting the total number of comparisons, it is not possible to properly evaluate the results.
- Ideally, all analyses should be planned before collecting data, and all should be reported.

Types of ANOVAs

Type	2-way ANOVA for independent groups			repeated measures ANOVA			mixed ANOVA		
Participants		Condition I	Condition II		Condition I	Condition II		Condition I	Condition II
	Task I	Participant group A	Participant group B	Task I	Participant group A	Participant group A	Task I	Participant group A	Participant group B
	Task II	Participant group C	Participant group D	Task II	Participant group A	Participant group A	Task II	Participant group A	Participant group B

Between-subject design

Within-subject design

both

NOTE

You may have more than 2 levels in each condition/task.

Kruskal-Wallis test - Non Parametric

- The Kruskal-Wallis test (Kruskal & Wallis, 1952) is the non-parametric counterpart of the one-way independent ANOVA (analysis of variance).
- The theory is very similar to that of the Mann-Whitney (and Wilcoxon rank-sum) test:
 - It is based on ranked data.
 - The sum of ranks for each group is denoted by R_i (where i is used to denote the particular group).

Kruskal-Wallis test Theory

- Once the sum of ranks has been calculated for each group.
- The test statistic, H , is calculated as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \quad (1)$$

- R_i is the sum of ranks for each group.
- N is the total sample size.
- n_i is the sample size of a particular group.

Example

- Open dataset youthcohort.dat.
 - Taken from Quantitative Data Analysis in Education, Paul Connolly.
 - Dataset Descriptor: <http://cw.routledge.com/textbooks/9780415372985/pdfs/youthcohort.pdf>
- Question:
 - "Are there any differences between pupils of different ethnicity in England and Wales in relation to the grades they achieved in GCSE Maths?"
 - H_0 : There is no difference.
 - H_A : There is a difference.
- Variables:
 - *ethsfr* ethnicity (nominal).
 - 1 White, 2 Black, 3 Indian, 4 Pakistani, 5 Bangladeshi, 6 Other Asian (inc Chinese), 7 Other ethnic groups (inc mixed), 8 Not answered, 9 Should be empty.
 - *gradmath* grade achieved in maths (ordinal).
 - -9 Didn't take 1 A* 2 A 3 B 4 C 5 D 6 E 7 F 8 G 9 Fail

Provisional Analysis - Statistics for gradmath grouped by ethnicity

In R:

```
library ( psych )  
describeBy ( as.numeric ( ydata$gradmath ) ,  
factor ( ydata$ethsfr ) )
```

Why *as.numeric* and *factor*? *describeBy* requires first argument to be a vector (sequence of datatypes, here we want to use the numeric values not the grade descriptors) and second to be a factor (set of integer values with associated descriptors).

In R

```
ydata=read.table("youthcohort.dat")  
Stats::kruskal.test(gradmath~ethsfr, data=ydata)
```

Output:

```
Kruskal-Wallis rank sum test data:  
gradmath by ethsfr Kruskal-Wallis  
chi-squared = 239.96, df = 6, p-value < 2.2e-16
```

In R

- Post-hoc test.
- We need the *as.numeric* to make sure our *gradmath* variable is treated as a numeric variable and we are using the numerical categories for *ethsfr* so we use the factors.

```
#Need library FSA to run the post-hoc tests  
library(FSA)  
tmp<-dunnTest(x=as.numeric(ydata$gradmath),  
g=factor(ydata$ethsfr), method="bonferroni")  
print(tmp, dunn.test.results = TRUE)
```

In R

$X = \text{gradmath}$, $g = \text{ethsfr}$

The test statistic is Z with significance is show underneath.

Col Mean- Row Mean	Banglade	Black	Indian	Other As	Other et	Pakistan
Black	-0.121395 1.0000					
Indian	7.537812 0.0000*	10.37020 0.0000*				
other As	7.104755 0.0000*	8.751802 0.0000*	0.969281 1.0000			
other et	4.680013 0.0001*	6.365208 0.0000*	-3.796034 0.0031*	-3.772717 0.0034*		
Pakistan	-0.971173 1.0000	-1.117740 1.0000	-11.64258 0.0000*	-9.688934 0.0000*	-7.549525 0.0000*	
white	4.292492 0.0004*	6.870085 0.0000*	-7.611379 0.0000*	-5.752329 0.0000*	-1.868248 1.0000	8.515972 0.0000*

alpha = 0.05

Reject H_0 if $p \leq \alpha$

Friedman's ANOVA

- Non-parametric test for differences between several related groups.
- Used for testing differences between conditions when:
 - There are more than two conditions.
 - The same participants have been used in all conditions (each case contributes several scores to the data).
- If you have violated some assumption of parametric tests then this test can be a useful way around the problem.

Theory of Friedman's ANOVA

- The theory for Friedman's ANOVA is much the same as the other tests: it is based on ranked data.
- Once the sum of ranks has been calculated for each group, the test statistic, F_r , is calculated as:

$$F_r = \left[\frac{12}{Nk(k+1)} \sum_{i=1}^k R_i^2 \right] - 3N(k+1) \quad (2)$$

Example

- Does the 'andikins' diet work?
- Dataset diet.dat (Andy Field)
- Variables:
 - Outcome: weight (Kg)
 - Time since beginning the diet
 - Baseline (start)
 - 1 Month
 - 2 Months
- Participants:
 - 10 people

Data for the diet sample with ranks

	Weight			Weight		
	Start	Month 1	Month 2	Start (Rank)	Month 1 (Rank)	Month 2 (Rank)
Person 1	63.75	65.38	81.34	1	2	3
Person 2	62.98	66.24	69.31	1	2	3
Person 3	65.98	67.7	77.89	1	2	3
Person 4	107.27	102.72	91.33	3	2	1
Person 5	66.58	69.45	72.87	1	2	3
Person 6	120.46	119.96	114.26	3	2	1
Person 7	62.01	66.09	68.01	1	2	3
Person 8	71.87	73.62	55.43	2	3	1
Person 9	83.01	75.81	71.63	3	2	1
Person 10	76.62	67.66	68.6	3	1	2

In R

```
#read in the data  
diet <- read.table("diet.dat")  
#treat our data as a matrix  
Responses <- na.omit(with(diet ,  
cbind(month1, month2, start)))  
cat("\\nMedians:\\n")#display the heading Medians  
#print out the medians of each of our variables  
print(apply(Responses , 2, median))  
#conduct the friedman test  
friedman.test(Responses)
```

In R

Output:

```
month1 month2 start  
68.575 72.250 69.225
```

```
Friedman rank sum test data: Responses  
Friedman chi-squared = 0.2, df = 2,  
p-value = 0.9048
```

Writing and interpreting the results

For Friedman's ANOVA we need only report the test statistic (X^2), its degrees of freedom and its significance:

- The weight of participants did not significantly change over the two months of the diet, $X^2(2) = 0.20, p = .91$.

15min break

Lab Exercise

- Using survey.dat
 - Are there differences in total perceived stress scores (tpstress) for participants with different education levels (educ)?
 - Are there differences in total self-esteem scores (tslfset) for participants for young, middle aged and old participants (agegp3)?
- Using youthcohort.dat
 - Are there differences in GCSE English grades (gradeng) achieved by students of different ethnicities (ethsfr)?
 - Are there differences in GCSE Maths grades (gradmath) achieved by students with parents of different educational attainment (s1pared)?

Comparing Nominal Values

Comparing Nominal Variables

When two variables are independent, there is no relationship between them.

- For all cases, the classification of a case into a particular category of one variable (the group variable) has no effect on the probability that the case will fall into any particular category of the second variable (the test variable).

Comparing Nominal Values

Analyzing categorical variables:

- The mean of a categorical variable is meaningless.
- The numeric values you attach to different categories are arbitrary.
- The mean of those numeric values will depend on how many members each category has.
- Therefore, we analyse frequencies.

Comparing Nominal Variables

- Using the **bullying.dat** dataset (Paul Connolly)
- Dataset Descriptor: <http://cw.routledge.com/textbooks/9780415372985/pdfs/bullying.pdf>
- Contains the results of national school level survey, subset of this survey which deals with bullying.
- Question:
 - "Is there a difference between boys and girls in terms of whether they have been bullied at school?"
 - H_0 : There is no difference.
 - H_A : There is a difference.
- Variables:
 - 'rsex' 1-Male, 2-Female, 99-not answered;
 - 'ubullsch' – Q26. Have you yourself ever been bullied in school? 1- Yes, 2 –No, 99-not answered

Chi-square Test

- We are comparing observed values to expected values according to a specific hypothesis (null hypothesis).
- The null hypothesis for this test states that the proportions (the distribution across categories) are the same for all of the populations.
- Views the data as two (or more) separate samples representing the different populations being compared.
 - The same variable is measured for each sample by classifying individual subjects into categories of the variable.
 - The data are presented in a matrix with the different samples defining the rows and the categories of the variable defining the columns.
 - The data, called **observed frequencies**, simply show how many individuals from the sample are in each cell of the matrix.

Chi Square Test for Independence

We are comparing observed values to **expected values** according to a specific hypothesis (null hypothesis):

- The null hypothesis for this test states that the proportions (the distribution across categories) are the same for all of the populations.
- The null hypothesis is used to construct an idealized sample distribution of **expected frequencies** that describes how the sample would look if the data were in perfect agreement with the null hypothesis.

Chi-Square Test for Independence

- A chi-square statistic is computed to measure the amount of discrepancy between the ideal sample (expected frequencies from H_0) and the actual sample data (the observed frequencies).
- Output is a cross tabulation table.

In R

```
#import the library gmodels  
library(gmodels)
```

```
#Read in the data  
bully<-read.table("bullying.dat")
```

```
#Use the Crosstable function  
CrossTable(predictor, outcome, fisher = TRUE,  
chisq = TRUE, expected = TRUE)
```

```
gmodels::CrossTable(bully$rsex, bully$ubullsch,  
fisher = TRUE, chisq = TRUE, expected = TRUE,  
sresid = TRUE, format = "SPSS")
```

Cell Contents

Count
Expected Values
Chi-square contribution
Row Percent
Column Percent
Total Percent
Std Residual

If we are looking at 2x 2 table we look at Yates continuity correction

Statistics for All Table Factors

Pearson's Chi-squared test

Chi² = 2.543274 d.f. = 1 p = 0.1107649

Pearson's Chi-squared test with Yates' continuity correction

Chi² = 2.300502 d.f. = 1 p = 0.1293322

Fisher's Exact Test for Count Data

Sample estimate odds ratio: 0.7778361

Alternative hypothesis: true odds ratio is not equal to 1

p = 0.1189103

95% confidence interval: 0.5630033 1.071191

Alternative hypothesis: true odds ratio is less than 1

p = 0.06431698

95% confidence interval: 0 1.019774

Alternative hypothesis: true odds ratio is greater than 1

p = 0.9531974

95% confidence interval: 0.59206 Inf

Minimum expected frequency: 100.2197

Need to check this – assumption is that lowest expected frequency is 5 or more. If not then this test is not appropriate and should use Fisher's Exact Probability Test.

Total Observations in Table: 801

bully\$sex	bully\$bullsch	No	Yes	Row Total
Female		318	154	472
		328.220	143.780	
		0.318	0.726	
		67.373%	32.627%	58.926%
		57.092%	63.115%	
		39.700%	19.226%	
Male		-0.564	0.852	
		239	90	329
		228.780	100.220	
		0.457	1.042	
		72.644%	27.356%	41.074%
		42.908%	36.885%	
Column Total		29.838%	11.236%	
		0.676	-1.021	
		557	244	801
		69.538%	30.462%	

Make sure your %s add up to 100.

Yate's Continuity Correction

- Again this is to prevent us making a Type I error.
- When we are doing a Chi-square test we are assuming that the probability distribution of the binomial frequencies observed approximate the Chi-square distribution.
 - This is not quite correct.
 - Yate's correction adjusts Pearson's Chi-square by subtracting 0.5 from each of the differences between the observed values and the expected values, which will reduce the Chi-square statistic and its associated p-value.
 - Particularly important for small data (larger datasets this can be overcome).
 - May tend to overcorrect - make sure you reflect your research area's perspective (some would say Yates is unnecessary even for small data).

In R - More Simplistic

```
#Create your contingency table  
mytable<-xtabs(~ubullsch+rsex, data=bully)  
  
ctest<-chisq.test(mytable, correct=TRUE)  
#get Yates correction needed for 2x2 table  
  
ctest  
ctest$expected #expected frequencies  
ctest$observed #observed frequencies  
ctest$p.value
```

OUTPUT:

X-squared = 2.3005, **df** = 1, p-value = 0.1293

Chi Square Statistic

- Measures the amount of discrepancy between the ideal sample (expected frequencies from H_0) and the actual sample data (the observed frequencies = f_o).
 - A large discrepancy results in a large value for chi-square which indicates that the data do not fit the null hypothesis and the hypothesis does not hold.
 - We compare our test statistic to the Chi Square distribution (in the form of a table) and we can get the probability that any deviation in the observed from the expected is due to chance only.
- For one degree of freedom:
 - The critical value associated with $p = 0.05$ for Chi Square is 3.84.
 - The critical value associated with $p = 0.01$ it is 6.64.
 - *Chi-square values higher than this critical value are associated with a statistically low probability that H_0 holds.*

Chi-Square Test for Independence

Reporting the findings:

A Chi-Square test for independence (with Yates' Continuity Correction) indicated no significant association between gender and reported experience of bullying, $\chi^2(1, n = 801) = 2.30, p = .13$.

Repeated Measures Categorical Variables

- Use McNemar's test.
- Matched Samples or repeated measures (pre-test/post-test).
- Two variables one recorded at Time 1 and one recorded at Time 2 (after an intervention).
- Use for categorical variables with two response options.
- For 3 or more use Cochran's Q-test.

In R: `mcnemar.test`.

Repeated Measures Categorical Variables

You can try this by using one of Pallant's Datasets.

- `experim4ED.dat`
- Fields Time1 Clinical Depression, Time 2 Clinical Depression.
- You should get a conclusion that indicates there is no significant change in those diagnosed as clinically depressed after the program of treatment.

Back to Hypothesis Testing

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.

Power Analysis

- An important aspect of experimental design.
- Allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence.
 - Also allows us to determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints.
 - If the probability is unacceptably low, we would be wise to alter or abandon the experiment.

Power Analysis

- ① sample size.
- ② effect size.
- ③ significance level.
 - $P(\text{Type I error}) = \text{probability of finding an effect that is not there.}$
- ④ power.
 - $P(\text{Type II error}) = \text{probability of finding an effect that is there.}$

Given any three, we can determine the fourth.

Power Analysis

The `pwr` package implements power analysis as outlined by Cohen.

Function	power calculations for
<code>pwr.2p.test</code>	two proportions (equal n)
<code>pwr.2p2n.test</code>	two proportions (unequal n)
<code>pwr.anova.test</code>	balanced one way ANOVA
<code>pwr.chisq.test</code>	chi-square test
<code>pwr.f2.test</code>	general linear model
<code>pwr.p.test</code>	proportion (one sample)
<code>pwr.r.test</code>	correlation
<code>pwr.t.test</code>	t-tests (one sample, 2 sample, paired)
<code>pwr.t2n.test</code>	t-test (two samples with unequal n)

Power Analysis

- For a one-way analysis of variance use:

`pwr.anova.test(k = , n = , f = ,
sig.level = , power =)`

- where k is the number of groups and n is the common sample size in each group.
- For a one-way ANOVA effect size is measured by f where Cohen suggests that f values of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes respectively.

Online Calculators

- A range of online calculators are available at <http://statpages.info/#Power>.
- You can also use Gpower: A downloadable application - <http://www.gpower.hhu.de/>.

How can I increase Power?

Increase Alpha level:

- Changing alpha from 0.05 to 0.10 will increase your power (better chance of finding significant results).
- Downsides to increasing your alpha level?
 - This will increase the chance of Type I error.
- This is rarely acceptable in practice.
- Only really an option when working in a new area:
 - Researchers are unsure of how to measure a new variable.
 - Researchers are unaware of confounders to control for.

How can I increase Power?

Increase N:

- Sample size is directly used when calculating p-values.
- Including more subjects will increase your chance of finding statistically significant results.
- Downsides to increasing sample size?
 - More subjects means more time/money.
- More subjects is ALWAYS a better option if possible.

How can I increase Power?

Use fewer groups/variables (simpler designs):

- Related to sample size but different.
 - 'Use fewer groups' NOT 'Use less subjects'.
- \uparrow groups negatively effects your degrees of freedom.
 - Remember, df is calculated with # groups and # subjects
- Lots of variables, groups and interactions make it more difficult to find statistically significant differences.
 - The purpose of the Family-wise error rate is to make it harder to find significant results!
- Downsides to fewer groups/variables?
 - Sometimes you NEED to make several comparisons and test for interactions - unavoidable.

How can I increase Power?

Measure variables more accurately:

- If variables are poorly measured (sloppy work, broken equipment, outdated equipment, etc.) this increases measurement error.
- More measurement error decreases confidence in the result.
- More of an internal validity problem than statistical problem.
- Downsides to measuring more accurately?
 - None - if you can afford the best tools.

How can I increase Power?

Decrease subject variability:

- Subjects will have various characteristics that may also be correlated with your variables.
 - Gender, race/ethnicity, age, etc.
 - These variables can confound your results, making it harder to find statistically significant results.
 - When planning your sample (to enhance power), select subjects that are very similar to each other.
 - This is a reason why repeated measures tests and paired samples are more likely to have statistically significant results.
- Downside to decreasing subject variability?
 - Will decrease your external validity - generalisability.
 - If you only test women, your results do not apply to men.

How can I increase Power?

Increase magnitude of the mean difference:

- If your groups are not different enough, make them more different:
 - Compare a 'very' high group to a 'very' low group.
 - Sampling at the extremes, getting rid of the middle group.
- Downsides to using the extremes?
 - Similar to decreasing subject variability, this will decrease your external validity.

The Catch-22 of Power and P-values

- The larger your sample, the more likely you'll find statistically significant results.
 - Sometimes miniscule differences between groups or tiny correlations are 'significant'.
 - This becomes relevant once sample size grows to 100 150 subjects per group.
 - Once you approach 1000 subjects, it's hard not to find $p < 0.05$.
- Example from most highly cited paper in Psych, 2004.

Example

The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance

Douglas A. Gentile^{a,*}, Paul J. Lynch^b, Jennifer Ruh Linder^c, David A. Walsh^a

^a *National Institute on Media and the Family, 606 24th Avenue South, Suite 606, Minneapolis, MN 55454, USA*

^b *University of Oklahoma Medical School, USA*

^c *Linfield College, USA*

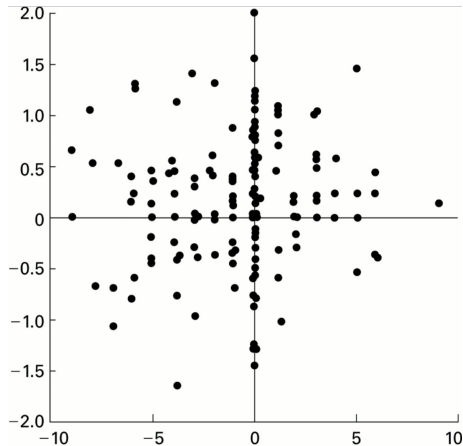
Table 4

Correlations between media habits and parental limits and outcomes ($n = 399-586$)

	Trait hostility	Arguments with teachers	Physical fights	Grades
Amount variables				
Amount of video game play	0.20***	0.12*	0.21***	-0.25***
Amount of time watching TV	0.20***	0.10*	0.12***	-0.20***
Amount of reading for pleasure	-0.08 [†]	-0.17***	-0.07	0.07 [†]

- This paper was the first to find a link between playing video games/TV and aggression in children.
- Every correlation in this table except 1 has $p < 0.05$.
- What does a correlation of 0.10 looks like?

Do you see a relationship between these two variables?



$$r = 0.1$$

Decisions

		Actual Situation	
		No Effect, H_0 True	Effect Exists, H_0 False
EXPERIMENTER'S DECISION	Reject H_0	Type I error	Decision correct
	Retain H_0	Decision correct	Type II error

Measuring Effect Size

- A hypothesis test evaluates the statistical significance of the results from a research study.
- The hypothesis test is influenced not only by the size of the effect/difference but also by the size of the sample.
- Thus, even a very small effect/difference can be significant if it is observed in a very large sample.
- Finding a statistically significant result does not necessarily mean a large effect.
- It is recommended that the hypothesis test be accompanied by a measure of the **effect size**.
- Cohen's measures of effect size are used as standard.

Effect Size

- To get an idea of how 'important' a difference or association is, we can use Effect Size.
 - There are over 40 different types of effect size.
 - Depends on statistical test used.
- Effect size is like a 'descriptive' statistic that tells you about the magnitude of the association or group difference.
 - Not impacted by statistical significance.
 - Effect size can stay the same even if p-value changes.
 - Present the two together when possible.

Another example

If we sample a group of 100 students and find their average IQ is 103.

- The population mean for IQ is 100, $SD = 15$.
- We run a one-sample t-test and find it to be statistically significant ($p < 0.05$)
- However, effect size is 0.2, or Small Effect.
- Interpretation: While this difference is likely not due to random sampling error - it's not very important either.