

Group Comparisons Explained

SCSRU Workshop

Statistical Consulting and Survey Research Unit
University of Waterloo

2026-10-07

Dataset Overview

Throughout this workshop, we will use the pokemon dataset.

```
str(df)
```

```
## 'data.frame':      801 obs. of  11 variables:
## $ name          : chr  "Bulbasaur" "Ivysaur" "Venusaur" "Charmeleon" ...
## $ weight_kg      : num  6.9 13 100 8.5 19 90.5 9 22.5 85.5 2.9 ...
## $ height_m       : num  0.7 1 2 0.6 1.1 1.7 0.5 1 1.6 0.3 ...
## $ hp             : int   45 60 80 39 58 78 44 59 79 45 ...
## $ attack         : int   49 62 100 52 64 104 48 63 103 30 ...
## $ defense        : int   49 63 123 43 58 78 65 80 120 35 ...
## $ sp_attack      : int   65 80 122 60 80 159 50 65 135 20 ...
## $ sp_defense     : int   65 80 120 50 65 115 64 80 115 20 ...
## $ speed          : int   45 60 80 65 80 100 43 58 78 45 ...
## $ type1          : chr   "grass" "grass" "grass" "fire" ...
## $ is_legendary   : int    0 0 0 0 0 0 0 0 0 0 ...
```

1. Review: Types of variables

It is common to see different variables in a data set. There are many types of variables, but we can generally classify the variables as:

- ▶ Discrete/categorical: a variable that can only take values over a finite set of values (or levels), e.g., `is_legendary`, `name`
- ▶ Continuous: a variable that can take any numerical value over a continuous range, e.g., `wright_kg`, `height_m`
- ▶ Ordinal: a variable that has a natural order, e.g., a likert scale from Strongly Agree to Strongly Disagree.

2. Review: Hypothesis Test

Oftentimes, we are interested to investigate the relationships between multiple variables. For example,

- ▶ Is the mean weight of pokemons 50kg?
- ▶ Are legendary pokemons on average higher than the non-legendary pokemons?
- ▶ How does pokemons' attack and defense change after they grow?

These questions lead us to *hypothesis testing*.

2.1 Steps of hypothesis testings

1. Formulate the null and alternate hypothesis.
 2. Choose and evaluate the appropriate test statistic (with R).
 3. Assess the strength of the evidence against the null hypothesis.
 4. Interpret the results.
- ▶ Note that Step (2) is a tedious (and sometimes iterative) process. It cannot be checked off using “a few clicks”.
 - ▶ Check section 9 for more detailed recap of hypothesis tests.

2.2 Caveat 1: Implication of a large p-value

- ▶ If $p\text{-value} \geq 0.05$, we do not have enough evidence (from the data) against H_0 . So, we failed to reject H_0 from the current sample/data.
- ▶ But this does NOT mean that we accept H_0 .
- ▶ A large p-value can be the result of 2 causes:
 - ▶ the sample is not general enough to represent the population (consider larger sample sizes, redesign the sampling process, or wait for more observations);
 - ▶ the test is not powerful enough to reveal patterns from the data (consider alternative tests).

2.3 Caveat 2: One-sided vs two-sided tests

- ▶ One-sided (`alternative=c('greater')` or `alternative=c('less')`): test if one group is on average greater or less than the other group.
 - ▶ When you are comparing two groups and you have a clear expectation that one group is greater than the other.
- ▶ Two-sided (`alternative=c('two-sided')`): test if one group is on average unequal to the other group.
 - ▶ When you are comparing multiple groups or when you do not have a clear idea of which group dominates the other.
 - ▶ You can always investigate pairwise relations among different groups via post-hoc tests.
- ▶ The two-sided test is usually the default option.

3. T-Tests

- ▶ One of the most commonly used statistical test is the t-test.
- ▶ The **one-sample t-test** is used to compare the mean of a variable to a hypothetical value. In most cases, the hypothetical value comes from theory.
- ▶ The **two-samples t-test** is used to compare the means of two variables. It is often used to determine whether a treatment has an effect on the population of interest, or whether two groups are different from one another.
- ▶ The **paired t-test** is commonly used to investigate the difference of a variable pre- and post-treatment. Oftentimes, every subject of the study produces a pair of observations, or two similar subjects will be paired up.

3.1 One-sample t-test

- ▶ For a one-sample t-test, the assumptions are:
 1. The population of the variable from which the sample is drawn from is **independently identically distributed** (IID) from Normal distribution.
 2. If the sample size is large enough, the assumption in (1) is not necessary because Central Limit Theorem applies.

3.1 Example: Is the average pokemon weight around 50kg?

- ▶ H_0 : Average pokemon weight is 50kg.
- ▶ H_A : Average pokemon weight is not 50kg.

```
t.test(df$weight_kg, mu=50)
```

One Sample t-test

```
data:  df$weight_kg
t = 2.9078, df = 780, p-value = 0.003744
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 53.69681 69.05940
sample estimates:
mean of x
 61.3781
```

3.1 Example: Is the average pokemon weight around 50kg?

- ▶ The p-value is less than 0.05 and hence, we have evidence against H_0 , i.e. the average pokemons' weight is not 50.
- ▶ Q: But is it higher or lower?

```
t.test(df$weight_kg, alternative="greater", mu=50)
```

One Sample t-test

```
data: df$weight_kg
```

```
t = 2.9078, df = 780, p-value = 0.001872
```

```
alternative hypothesis: true mean is greater than 50
```

```
95 percent confidence interval:
```

```
54.9341      Inf
```

```
sample estimates:
```

```
mean of x
```

```
61.3781
```

3.2 Two-samples t-test

- ▶ Pokemons can be categorized as legendary and non-legendary. Suppose we want to compare the weight/height/hp/... between these two groups.
- ▶ As in the one-sample t-test, there are assumptions for using the two-samples t-test:
 1. The population in which each sample was drawn from is **independently normally distributed**.
 2. The population variances are similar.
 - ▶ This is not as important when using R because the default in R is to assume unequal population variances. The results of tests for equal and unequal population variances will be the same if the population variances are the same.
 3. When sample size is large enough, (1) is not important since Central Limit Theorem applies.

3.2 Example: Is avg weight of legendary pokemons the same as the non-legendary?

- ▶ H_0 : The avg weight of legendary and non-legendary pokemons are the same.
- ▶ H_A : The avg weight of legendary and non-legendary pokemons are NOT the same.

```
t.test(df$weight_kg[df$is_legendary == 1],  
       df$weight_kg[df$is_legendary == 0])
```

Welch Two Sample t-test

```
data: df$weight_kg[df$is_legendary == 1] and df$weight_kg[df$is_legendary == 0]  
t = 5.1176, df = 69.188, p-value = 2.65e-06  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 92.34922 210.33790  
sample estimates:  
mean of x mean of y  
199.35072  48.00716
```

3.2 Example: comparing weights of pokemons

- ▶ H_0 : The avg weight of legendary and non-legendary pokemons are the same.
- ▶ H_A : The avg weight of legendary and non-legendary pokemons are NOT the same.

```
t.test(df$weight_kg[df$is_legendary == 1],  
       df$weight_kg[df$is_legendary == 0])
```

Welch Two Sample t-test

```
data: df$weight_kg[df$is_legendary == 1] and df$weight_kg[df$is_legendary == 0]  
t = 5.1176, df = 69.188, p-value = 2.65e-06  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 92.34922 210.33790  
sample estimates:  
mean of x mean of y  
199.35072  48.00716
```

3.2 Example: comparing heights of pokemons

- ▶ H_0 : The avg height of legendary pokemons are greater than that of non-legendary pokemons.
- ▶ H_A : The avg height of legendary pokemons are lower than that of non-legendary pokemons.

```
t.test(df$height_m[df$is_legendary == 1],  
       df$height_m[df$is_legendary == 0],  
       alternative = "less")
```

Welch Two Sample t-test

```
data: df$height_m[df$is_legendary == 1] and df$height_m[df$is_legendary == 0]  
t = 5.6684, df = 71.593, p-value = 1  
alternative hypothesis: true difference in means is less than 0  
95 percent confidence interval:  
 -Inf 1.58583  
sample estimates:  
mean of x mean of y  
 2.281159  1.055618
```

3.3 Paired t-test

- ▶ When the two samples are highly correlated, assumptions of two-sample t-test are violated.
- ▶ For example, pokemons' attack and defense before and after they evolve (supergrow) are highly correlated because different pokemon can have difference attack and defense abilities.

In another word, we may be interested in: - Are pokemons' attack on average unchanged after they evolve? - Are pokemons' defense on average higher after they evolve?

3.3 Example: comparing attacks of pokemons

- H_0 : Pokemons' attacks are the same before and after they evolve.

```
t.test(df$attack, df$sp_attack, paired = TRUE)
```

```
##  
## Paired t-test  
##  
## data: df$attack and df$sp_attack  
## t = 5.1137, df = 800, p-value = 3.956e-07  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
##  4.036861 9.066759  
## sample estimates:  
## mean difference  
##      6.55181
```

3.3 Example: comparing defenses of pokemons

```
t.test(df$defense, df$sp_defense,  
       alternative = "less", paired = TRUE)
```

```
##  
## Paired t-test  
##  
## data: df$defense and df$sp_defense  
## t = 2.0698, df = 800, p-value = 0.9806  
## alternative hypothesis: true mean difference is less than 0  
## 95 percent confidence interval:  
##      -Inf 3.766043  
## sample estimates:  
## mean difference  
##      2.097378
```

4. Analysis of variance (ANOVA)

- ▶ The name analysis of variance can be misleading. It is actually a test on means over more than 2 samples.
- ▶ Remember that we have t-test to test on means of 1 or 2 samples.
- ▶ In one-way ANOVA, the pooled variance two sample t-test is extended to more than two samples.
- ▶ Pokemons have different types:

```
## tp
##      bug      fire  grass normal  other  water
##      72      52      78    105    380    114
```

- ▶ We can use One-way ANOVA to test whether different types of pokemons have similar weight/height/hp/attack/defense
...

4.1 ANOVA assumptions

1. The samples are independent.
2. The errors are normally distributed.
3. The errors have equal variances. ANOVA works poorly if the variances are extremely different.

4.2 One-way ANOVA

In one-way ANOVA, we test the null hypothesis that k populations all have the same mean

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

against the alternative hypothesis that the population means are not all equal.

Example 5.2: Pokemons' hp over different types

- H_0 : Different types of pokemons have the same level of hp.

```
aov_model = aov(hp~factor(type1), data=df)
summary(aov_model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor(type1)	17	28555	1679.7	2.452	0.000936	***
Residuals	783	536473	685.2			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Note: Before making any interpretation, please ensure model fit is appropriate.

4.3 Two-way ANOVA

- ▶ In reality, we often want to understand the impact of two independent variables and their combination on the response variable.
- ▶ We can compare the hp/weight/height of pokemons that are of different types and are either legendary or non-legendary.
- ▶ Instead of performing two individual t-tests or one-way ANOVA tests, we could perform a two-way ANOVA.
- ▶ In general, when the response variables of the one-way ANOVA are the same, we try to use one model.
- ▶ As mentioned earlier, performing multiple statistical tests is not only tedious, but also increases the risk of a Type I error.

Example 4.3: Pokemons' attack over different types and whether they are legendary

- H_0 : Pokemons' attack are indifferent over different types and whether they are legendary

```
aov_model = aov(hp~factor(type1)*factor(is_legendary), data=df)
summary(aov_model)
```

```

              Df Sum Sq Mean Sq F value    Pr(>F)
factor(type1)  17  28555    1680    2.751 0.000183 ***
factor(is_legendary)  1  54154   54154  88.692 < 2e-16 ***
factor(type1):factor(is_legendary)  15  13998     933   1.528 0.088934 .
Residuals        767 468321     611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4 Checking ANOVA assumptions

- ▶ Suppose we are interested in the association of pokemons' attack with pokemon type.

```
model = aov(attack ~ type1, data = df)
summary(model)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## type1         17  85994     5058   5.343 2.37e-11 ***
## Residuals    783 741358       947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

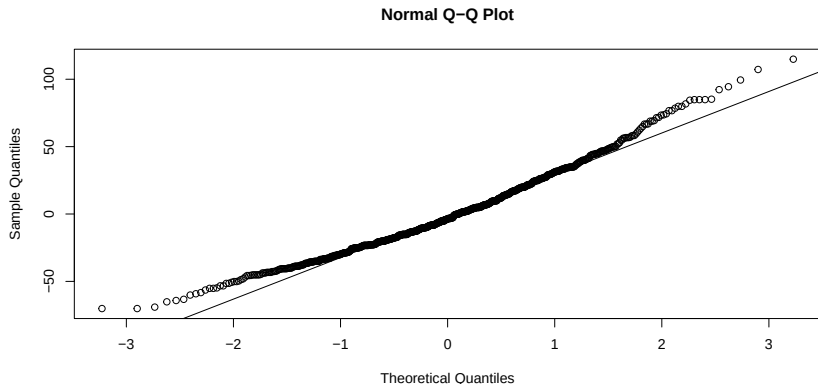
4.4 Checking ANOVA assumptions: Independence

- ▶ Checking the samples are independent is more by the design of the study and the sampling process.
- ▶ In the pokemon example, pokemons are independently captured and analyzed, so let's believe they are independent from each other.

4.4 Checking ANOVA assumptions: Normality

- ▶ Normal Quantile-Quantile (QQ) plot.
- ▶ Check if the dots stay aligned with the straight line.

```
qqnorm(residuals(model)); qqline(residuals(model))
```



4.4 Checking ANOVA assumptions: Normality

- ▶ Shapiro-Wilk test.
- ▶ Null hypothesis is that the model residuals are follow Normal distribution.

```
shapiro.test(residuals(model))
```

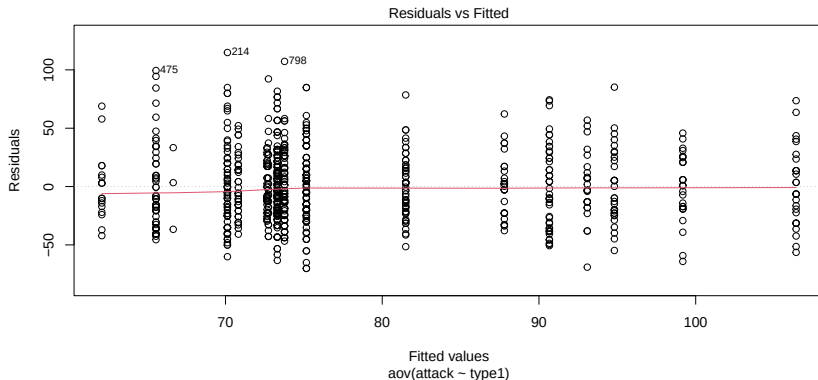
```
##  
##  Shapiro-Wilk normality test  
##  
## data:  residuals(model)  
## W = 0.97789, p-value = 1.18e-09
```

- ▶ Q: What can you say from the test result?

4.4 Checking ANOVA assumptions: Equal variance

- Residual vs fitted value plot.

```
plot(model, which=1)
```



4.4 Checking ANOVA assumptions: Equal variance

- ▶ Check if the dots
 - ▶ have no pattern or trend,
 - ▶ evenly centered around zero,
 - ▶ no significant variation in the span of values.
- ▶ Check if the red curve is almost straight and horizontal.
- ▶ There are also the Bartlett's and Levene's tests.

4.5 Difference between ANOVA and linear regression (LR)

1. Purpose

- ▶ ANOVA: test for statistically significant differences between the means of different groups
- ▶ LR: predict a continuous dependent variable based on several independent variables

2. Focus of analysis

- ▶ ANOVA: comparing group means, no prediction
- ▶ LR: quantifying the linear relationships and predicting outcomes

3. Variable types

- ▶ ANOVA: categorical independent variable and continuous dependent variable
- ▶ LR: continuous or categorical independent variable and continuous dependent variable

4.5 Difference between ANOVA and linear regression (LR)

4. Assumptions:

- ▶ ANOVA: Normality, homoscedasticity, independence
- ▶ LR: Normality, homoscedasticity, independence, linearity

5. Extension:

- ▶ ANOVA: One-way, Two-way ANOVA, MANOVA
- ▶ LR: Multiple regression, logistic regression, mixed models

6. Equivalent Model:

- ▶ ANOVA: LR with only categorical variables
- ▶ LR: Generalized linear models

5. Post-hoc tests

- ▶ Why post-hoc?
 - ▶ From ANOVA test, we know that whether the means (hp/height/weight/attack) are the same in different groups (type of pokemons).
 - ▶ But how do they differ from each other? Which group has higher mean than they other?
- ▶ Recall the ANOVA model that we use to examine the attack of pokemons of different types.

```
model = aov(attack ~ type1, data = df)
```

5.1 Tukey HSD

- ▶ Compare all possible pairs of means, good balance of test power and error control.
- ▶ Good for balanced design (equal group sizes), but is reasonably robust against unequal group sizes.
- ▶ Good default choice if you have a balanced design and care about all pairs.

```
TukeyHSD(model)$type1
```

##	diff	lwr	upr	p adj
## dark-bug	17.668103448	-6.02169655	41.35790344	4.467854e-01
## dragon-bug	36.282407407	11.97518015	60.58963467	3.306170e-05
## electric-bug	0.695512821	-20.72002039	22.11104603	1.000000e+00
## fairy-bug	-8.013888889	-36.39861503	20.37083725	9.999510e-01
## fighting-bug	29.053571429	5.06409937	53.04304349	3.297412e-03
## fire-bug	11.375000000	-8.22738330	30.97738330	8.532895e-01
## flying-bug	-3.458333333	-66.92851050	60.01184383	1.000000e+00
## ghost-bug	2.615740741	-21.69148652	26.92296800	1.000000e+00
## grass-bug	3.644230769	-13.95922909	21.24769063	9.999995e-01
## ground-bug	24.687500000	1.80300218	47.57199782	1.969708e-02
## ice-bug	3.179347826	-22.61933842	28.97803407	1.000000e+00
## normal-bug	5.036904762	-11.44440378	21.51821331	9.998532e-01
## poison-bug	2.531250000	-20.35324782	25.41574782	1.000000e+00
## psychic-bug	-4.558962264	-24.05367221	14.93574768	9.999968e-01
## rock-bug	20.541666667	0.07314956	41.01018377	4.804258e-02
## steel-bug	22.958333333	-2.42973753	48.34640420	1.333948e-01
## water-bug	3.182017544	-13.03248961	19.39652470	9.999998e-01

5.2 Bonferroni correction

- ▶ A very conservative test, control Type I error (low false positives); but could be too strict and miss true differences.
- ▶ Use if you have lots of comparisons and high need to avoid false positives (e.g., clinical studies).

```
pairwise.t.test(pokemon$attack, pokemon$type1,  
                p.adjust.method = "bonferroni")
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data:  pokemon$attack and pokemon$type1  
##  
##          bug      dark      dragon  electric  fairy  fighting  fire  flying  
## dark      1.00000 -          -          -          -          -          -  
## dragon    3.4e-05 1.00000 -          -          -          -          -  
## electric  1.00000 1.00000 0.00069 -          -          -          -  
## fairy     1.00000 0.84771 0.00041 1.00000 -          -          -  
## fighting  0.00384 1.00000 1.00000 0.03258 0.01117 -          -  
## fire      1.00000 1.00000 0.10356 1.00000 1.00000 1.00000 -          -  
## flying    1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 -  
## ghost     1.00000 1.00000 0.00976 1.00000 1.00000 0.22988 1.00000 1.00000  
## grass     1.00000 1.00000 0.00037 1.00000 1.00000 0.02924 1.00000 1.00000  
## ground    0.02620 1.00000 1.00000 0.17236 0.05038 1.00000 1.00000 1.00000  
## ice       1.00000 1.00000 0.02468 1.00000 1.00000 0.44295 1.00000 1.00000  
## normal    1.00000 1.00000 0.00046 1.00000 1.00000 0.03968 1.00000 1.00000  
## poison    1.00000 1.00000 0.00460 1.00000 1.00000 0.13867 1.00000 1.00000  
## psychic   1.00000 0.27984 4.2e-06 1.00000 1.00000 0.00053 1.00000 1.00000
```

5.3 Fisher LSD

- ▶ High sensitivity and test power, but also high chance of false positives (especially with many groups).
- ▶ Good if you have fewer groups and want maximum sensitivity.
- ▶ Q: Check how p-values differ from Bonferroni correction.

```
pairwise.t.test(pokemon$attack, pokemon$type1,  
                p.adjust.method = "none")
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data:  pokemon$attack and pokemon$type1  
##  
##          bug      dark      dragon  electric  fairy   fighting  fire    flying  
## dark      0.00921 -          -          -          -          -          -          -  
## dragon    2.2e-07 0.02397 -          -          -          -          -          -  
## electric  0.90952 0.02476 4.5e-06 -          -          -          -          -  
## fairy     0.32331 0.00554 2.7e-06 0.32086 -          -          -          -  
## fighting  2.5e-05 0.16294 0.38402 0.00021 7.3e-05 -          -          -  
## fire      0.04256 0.37780 0.00068 0.10173 0.02148 0.01446 -          -  
## flying    0.84878 0.25795 0.03413 0.82179 0.81240 0.08238 0.41711 -  
## ghost     0.70650 0.06775 6.4e-05 0.80322 0.25661 0.00150 0.23048 0.74575  
## grass     0.46887 0.03645 2.4e-06 0.62524 0.14776 0.00019 0.16091 0.69492  
## ground    0.00017 0.37386 0.14970 0.00113 0.00033 0.58363 0.05452 0.13020  
## ice       0.66630 0.09212 0.00016 0.75890 0.24806 0.00290 0.28783 0.72537  
## normal    0.28504 0.05072 3.0e-06 0.45204 0.09680 0.00026 0.22484 0.63741  
## poison    0.69872 0.05538 3.0e-05 0.80256 0.24511 0.00091 0.20120 0.74725  
## psychic   0.41325 0.00183 2.8e-08 0.41852 0.68076 3.5e-06 0.00814 0.95195
```

5.4 Scheffé's test

- ▶ A very conservative version of TukeyHSD, safe against false positives, works with unequal sample sizes.
- ▶ Use if you have very unbalanced groups, want to avoid false positive, care less of test power.

5.4 Scheffé's test (output)

```
##
## Posthoc multiple comparisons of means: Scheffe Test
## 95% family-wise confidence level
##
## $type1
##          diff      lwr.ci      upr.ci    pval
## dark-bug      17.668103448 -18.0198710  53.356078 0.9854
## dragon-bug     36.282407407  -0.3357027  72.900518 0.0567
## electric-bug    0.695512821 -31.5663468  32.957372 1.0000
## fairy-bug     -8.013888889 -50.7746290  34.746851 1.0000
## fighting-bug   29.053571429  -7.0858500  65.192993 0.3927
## fire-bug       11.375000000 -18.1554036  40.905404 0.9993
## flying-bug     -3.458333333 -99.0742550  92.157588 1.0000
## ghost-bug       2.615740741 -34.0023694  39.233851 1.0000
## grass-bug       3.644230769 -22.8748545  30.163316 1.0000
## ground-bug     24.687500000  -9.7873108  59.162311 0.6481
## ice-bug         3.179347826 -35.6856006  42.044296 1.0000
## normal-bug      5.036904762 -19.7916931  29.865503 1.0000
## poison-bug      2.531250000 -31.9435608  37.006061 1.0000
## psychic-bug    -4.558962264 -33.9271591  24.809235 1.0000
## rock-bug       20.541666667 -10.2935415  51.376875 0.7777
## steel-bug      22.958333333 -15.2880353  61.204702 0.9014
## water-bug       3.182017544 -21.2446520  27.608687 1.0000
## dragon-dark    18.614303959 -24.7806860  62.009294 0.9973
## electric-dark  -16.972590628 -56.7603620  22.815181 0.9975
## fairy-dark     -25.681992337 -74.3720330  23.008048 0.9712
## fighting-dark  11.385467980 -31.6063562  54.377292 1.0000
## fire-dark      -6.293103448 -43.9000695  31.313863 1.0000
## flying-dark    -21.126436782 -119.5370238  77.284150 1.0000
## ghost-dark     -15.052362708 -58.4473526  28.342627 0.9998
## grass-dark     -14.023872679 -49.3155295  21.267784 0.9990
## ground-dark     7.019396552 -34.5829080  48.621701 1.0000
## ice-dark       -14.488755622 -59.7957460  30.818235 1.0000
## normal-dark    -12.631198686 -46.6708584  21.408461 0.9996
```

5.5 Dunnett's test

- ▶ Focus on many-to-one comparisons; specifically useful if you have several treatment and are interested in how each treatment group differs from the control group.

5.5 Dunnett's test (result)

```
##
##   Dunnett's test for comparing several treatments with a control :
##     95% family-wise confidence level
##
## $bug
##           diff      lwr.ci   upr.ci    pval
## dark-bug      17.6681034 -2.156469 37.49268 0.12053
## dragon-bug     36.2824074 15.941148 56.62367 5.4e-06 ***
## electric-bug   0.6955128 -17.225862 18.61689 1.00000
## fairy-bug     -8.0138889 -31.767363 15.73959 0.99432
## fighting-bug  29.0535714  8.978222 49.12892 0.00041 ***
## fire-bug      11.3750000 -5.029058 27.77906 0.42320
## flying-bug    -3.4583333 -56.572716 49.65605 1.00000
## ghost-bug      2.6157407 -17.725519 22.95700 1.00000
## grass-bug      3.6442308 -11.087048 18.37551 0.99985
## ground-bug    24.6875000  5.536837 43.83816 0.00274 **
## ice-bug        3.1793478 -18.410025 24.76872 1.00000
## normal-bug     5.0369048 -8.755313 18.82912 0.98746
## poison-bug     2.5312500 -16.619413 21.68191 1.00000
## psychic-bug   -4.5589623 -20.872915 11.75499 0.99932
## rock-bug      20.5416667  3.412793 37.67054 0.00743 **
## steel-bug     22.9583333  1.712580 44.20409 0.02407 *
## water-bug      3.1820175 -10.386930 16.75097 0.99993
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Final Remarks

- ▶ So far we have introduced several tests and ways to examine statistical significance.
- ▶ But we cannot cover everything that you need to be aware of when you start to perform a formal statistical analysis.
- ▶ We would like to emphasize some key caveats that would help you avoid drawing incorrect conclusions from your analysis results.

6.1 Caveat 1: Statistical significance vs practical significance

- ▶ Statistical inference techniques test for statistical significance.
- ▶ Statistical significance means that the effect observed in a sample is very unlikely to occur if the null hypothesis is true.
- ▶ Whether this observed effect has practical importance is an entirely different question. The experts in the field of interest determine whether these results have any practical importance.

6.2 Caveat 2: Danger of over reliance on p-values

The ASA's Statement on p-values:

- ▶ P-values can indicate how incompatible the data are with a specified statistical model.
- ▶ P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- ▶ Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- ▶ Proper inference requires full reporting and transparency
- ▶ A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- ▶ By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

6.3 Caveat 3: Model diagnostics

- ▶ The statistical test results are only valid when the test assumptions are satisfied by the data.
- ▶ We did not check model assumptions when we performed the tests above, but this is a step that we cannot skip when doing actual analysis.
- ▶ There are a variety of model diagnostics test for different model assumptions.
 - ▶ **quantile-quantile plot** (QQ-plot). When the assumptions of residuals normality is met, we expect the points to lie on a straight line.
 - ▶ **residuals against the explanatory variables**. When the assumption of independent error is met, we expect the points to scatter randomly around the the horizontal line $y = 0$.

6.4 Caveat 4: Model fitting procedure

The MIDI steps of data analysis

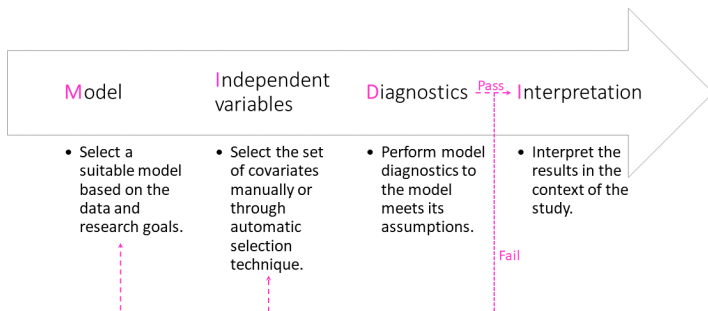


Figure 1: Recommended steps to data analysis

Next steps

Now that you have reviewed how to perform some common statistics with R, you can learn more about the different models, or explore other models such as

- ▶ exploratory data analysis,
- ▶ linear regression with R,
- ▶ generalized linear models,
- ▶ count data analysis, etc.

The SCSRU organize similar workshops to this on a regular basis to improve quality of research and data literacy among the UWaterloo community. We also provide 1-1 free consultation to all researchers on campus. More information are available on our website.

Thank you!

The Statistical Consulting and Survey Research Unit (SCSRU) is the unit through which the Department of Statistics and Actuarial Science provides statistical advice to those working on research problems.