# Statistics (part 1)

Ting Xu

# Statistics

## Two concepts and Two Theorems

Sample & Population

Theorem 1: Law of Large Numbers

Theorem 2: Central Limit Theorem

## Important Distributions

Normal (mu, sigma2)

Chi-square (df)

T(df)

F(df1, df2)

## Hypothesis testing and Two types of errors

Null hypothesis (H0) & Alternative hypothesis (H1)

False Positive & False Negative Errors

## Statistical tests

One-sample t test

Two-sample t-test

Paired t-test

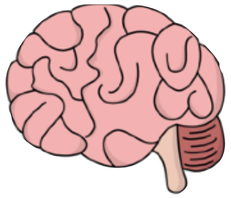
ANOVA

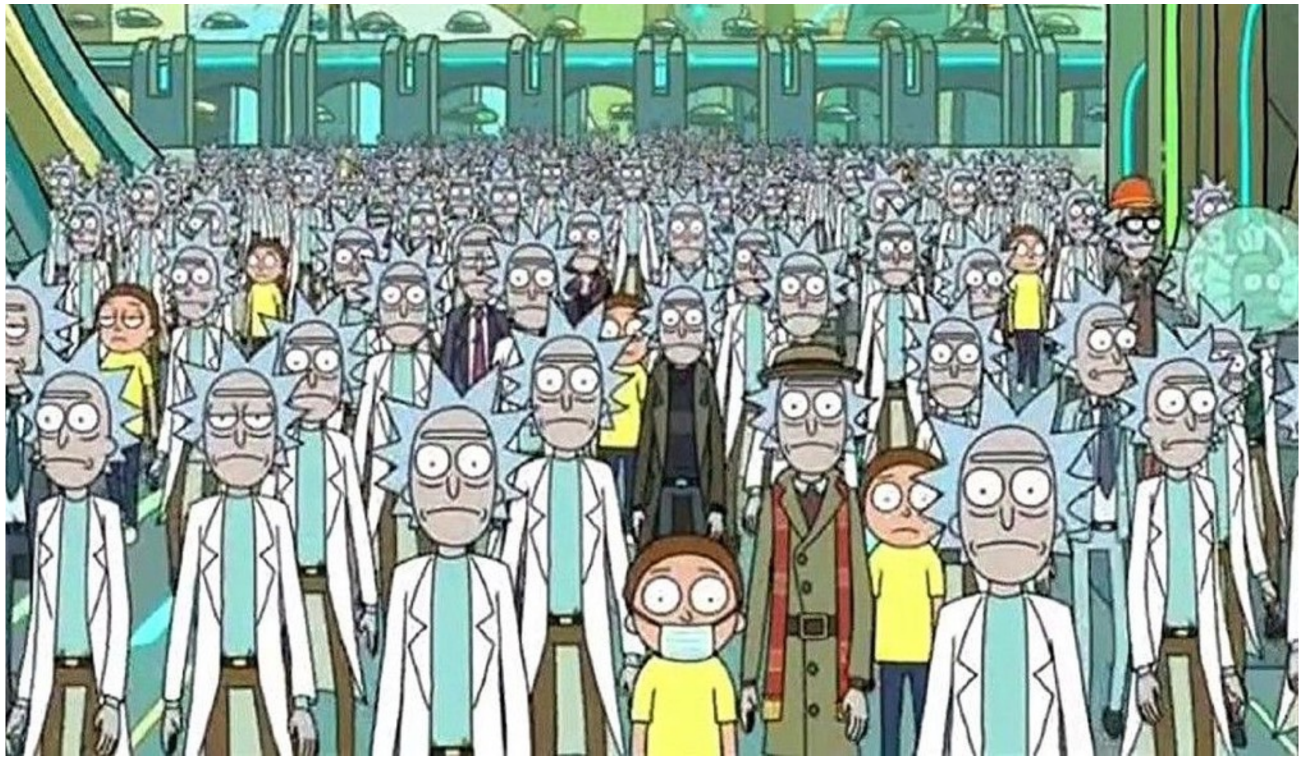
Repeated measure ANOVA

Regression

**Sample:** A fraction

**Population:** A whole

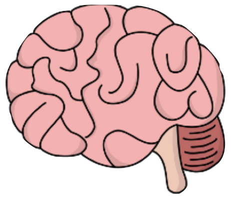


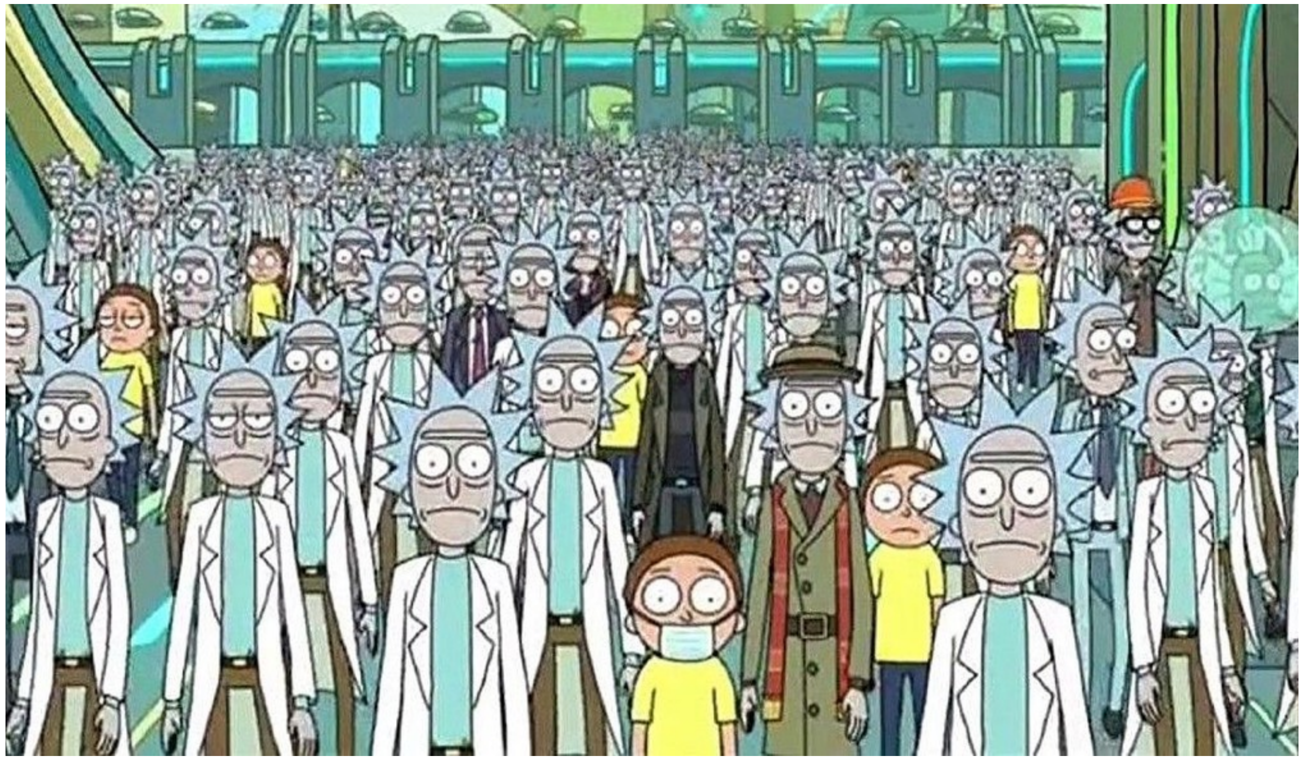


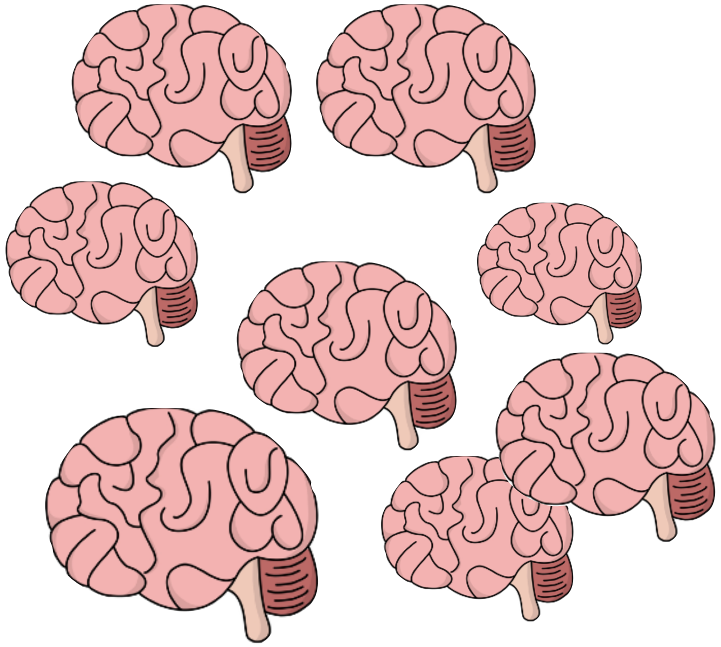
**Theorem 1: Law of Large Numbers**: the sample averages converge almost surely (converge in probability) to the expected value. In other words, as a sample size grows, its average gets closer to the average of the whole population.

**Sample:** A fraction

**Population:** A whole





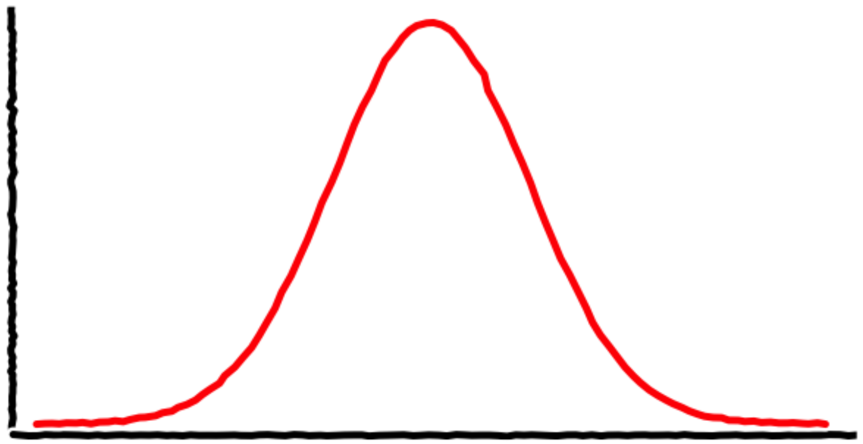


Sample (1) -> mean

Sample (2) -> mean

...

Sample (n) -> mean



NORMAL DISTRIBUTION

**Theorem 1: Law of Large Numbers**: the sample averages converge almost surely (converge in probability) to the expected value. In other words, as a sample size grows, its average gets closer to the average of the whole population.

**Theorem 2: Central Limit Theorem**: when independent random variables are summed up, their properly normalized sum tends toward a normal distribution, even if the original variables themselves are not normally distributed. In other words, as the sample size gets larger, the distribution of sample means approximates a normal distribution, regardless of the population's distribution.

# What is the Hypothesis?

An assertion (conjecture) concerning one or more **populations**

A medication has an effect
Male's brain is larger than female

**H0 (Null Hypothesis)**
**H1 (Alternative Hypothesis)**

H0: Medication effect=0
H1: Medication effect>0

H0: Brain size: Male=Female
H1: Brain size: Male>Female

**Alpha:** the probability of rejecting the null hypothesis when it is true

# What is the Hypothesis?

An assertion (conjecture) concerning one or more **populations**

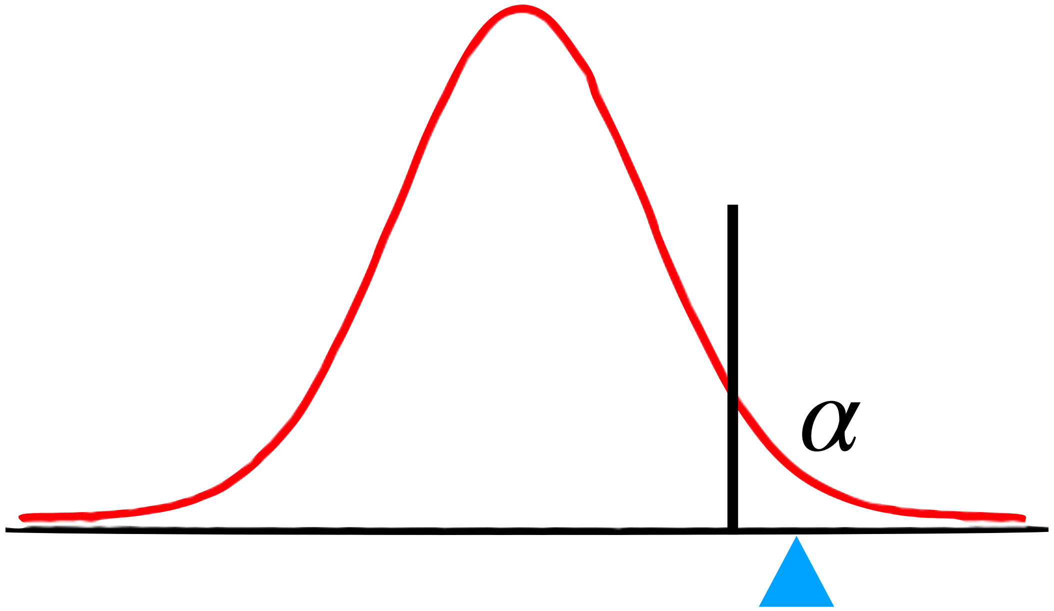
A medication has an effect
Male's brain is larger than female

**H0 (Null Hypothesis)**
**H1 (Alternative Hypothesis)**

H0: Medication effect=0
<u>H1: Medication effect>0</u>

H0: Brain size: Male=Female
<u>H1: Brain size: Male>Female</u>

H0: $\mu$(population)=0
<u>H1: $\mu$(population)>0</u>




Observation: mean(sample)

# Type I and Type II Error

## Type I Error (false-positive)


$\alpha$ ***(p-value)***

## Type II Error (false-negative)


$\beta$

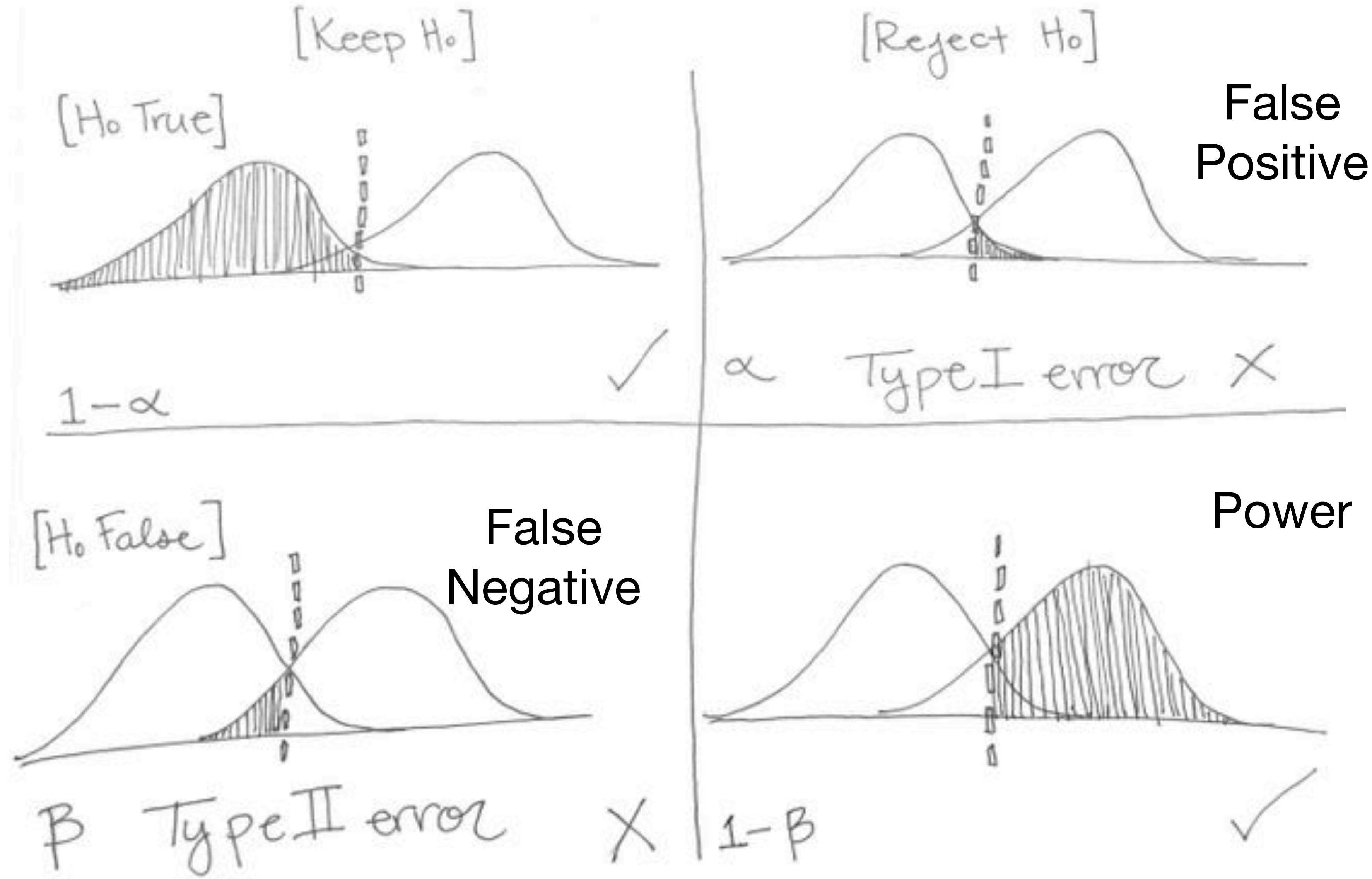
**Decision based on sample**

**Accept H0** | **Reject H0**

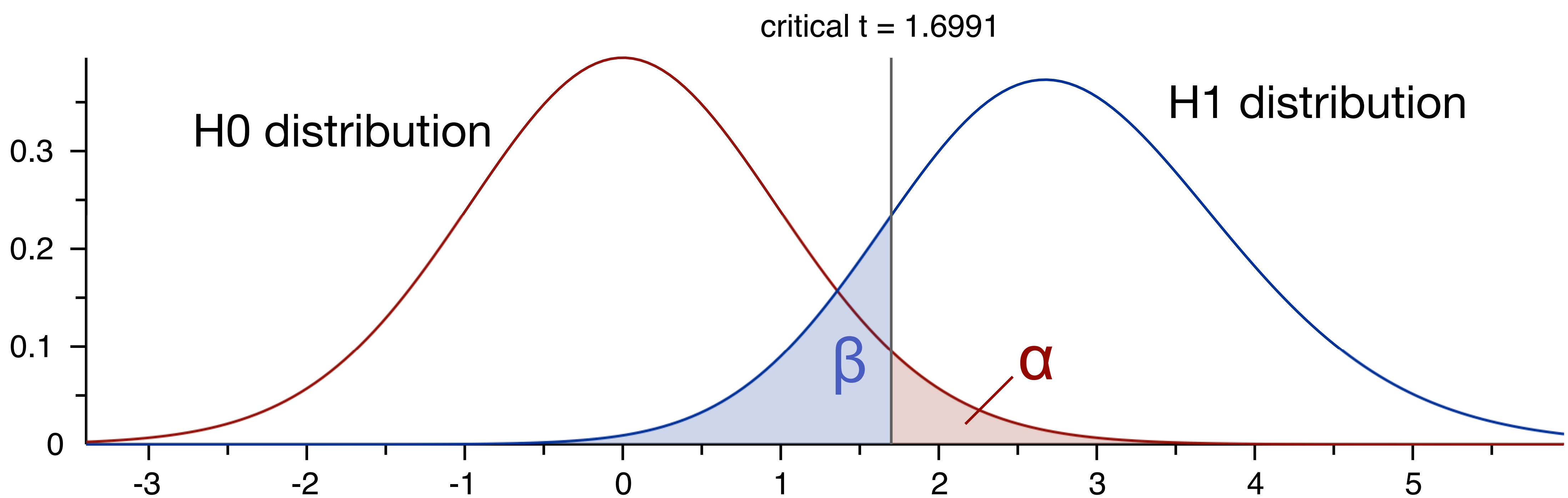
**Truth about the population**

**H0 is true**

**H1 is true**




**Alpha:** the probability of rejecting the null hypothesis when it is true (risk of type-I error)
**Power (1-beta):** the probability of rejecting the null correctly




```
t tests - Means: Difference from constant (one sample t-test)

Input:  One Tail(s)
        Effect size d                    =0.5
        α err prob                       =0.05
        Total sample size                =30
Output:
        Critical t                       =1.6991270
        Df                               =29
        Power (1-β err prob)             =0.8482542
```

# Statistics

## Two concepts and Two Theorems

Sample & Population

Theorem 1: Law of Large Numbers

Theorem 2: Central Limit Theorem

## Hypothesis testing and Two types of errors

Null hypothesis (H0) & Alternative hypothesis (H1)

False Positive & False Negative Errors

## Important Distributions

Normal (mu, sigma2)

Chi-square (df)

T(df)

F(df1, df2)

## Statistical tests

One-sample t test

Two-sample t-test

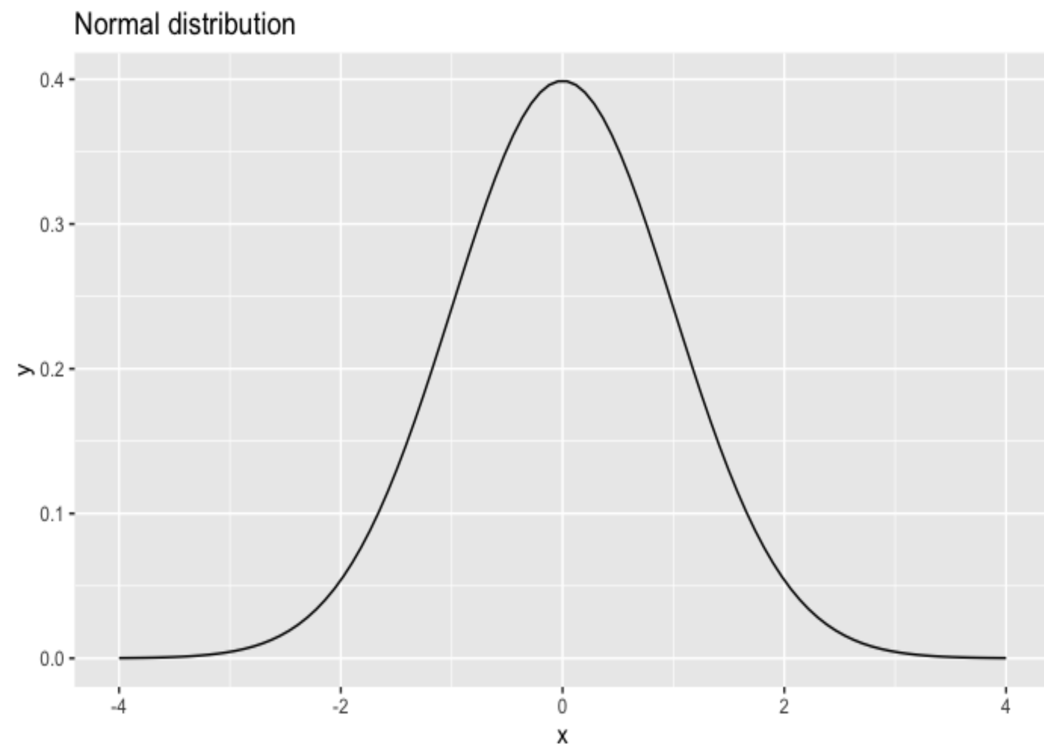
Paired t-test

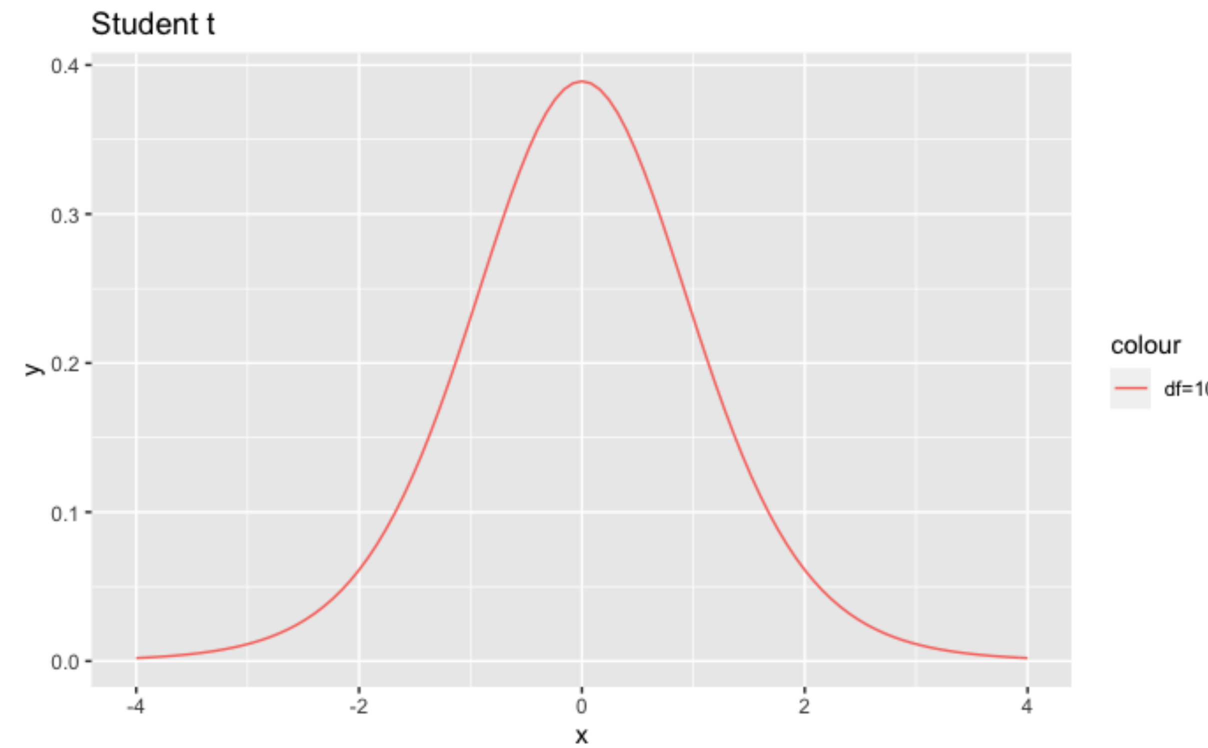
ANOVA

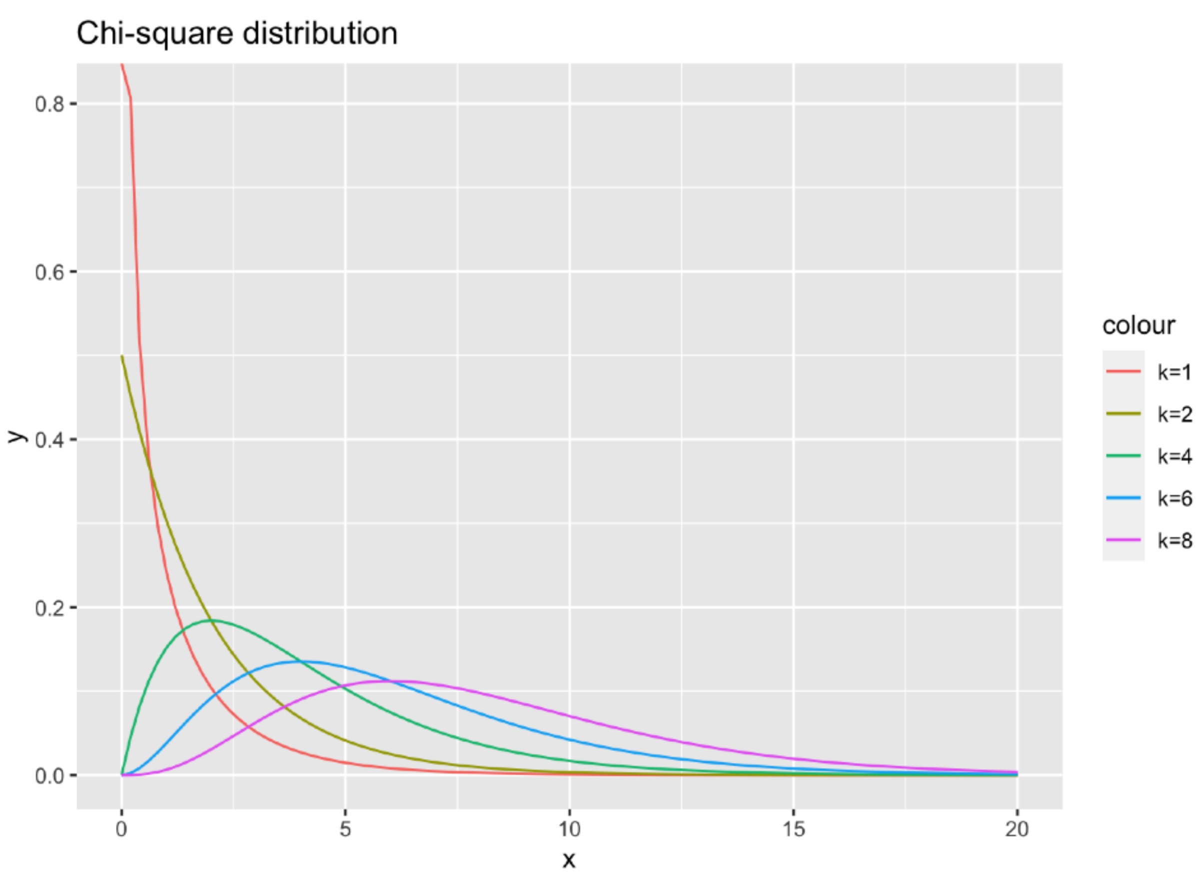
Repeated measure ANOVA

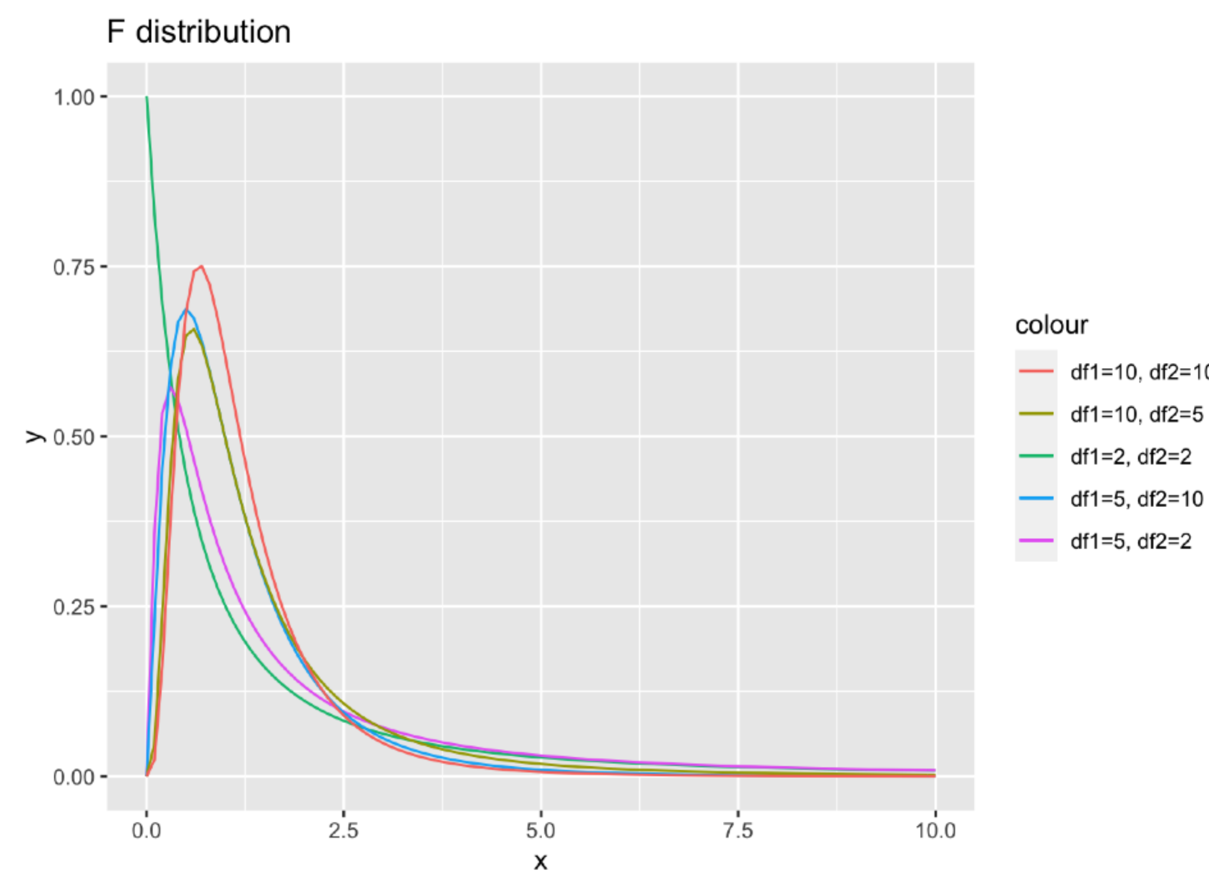
Regression

# Distributions

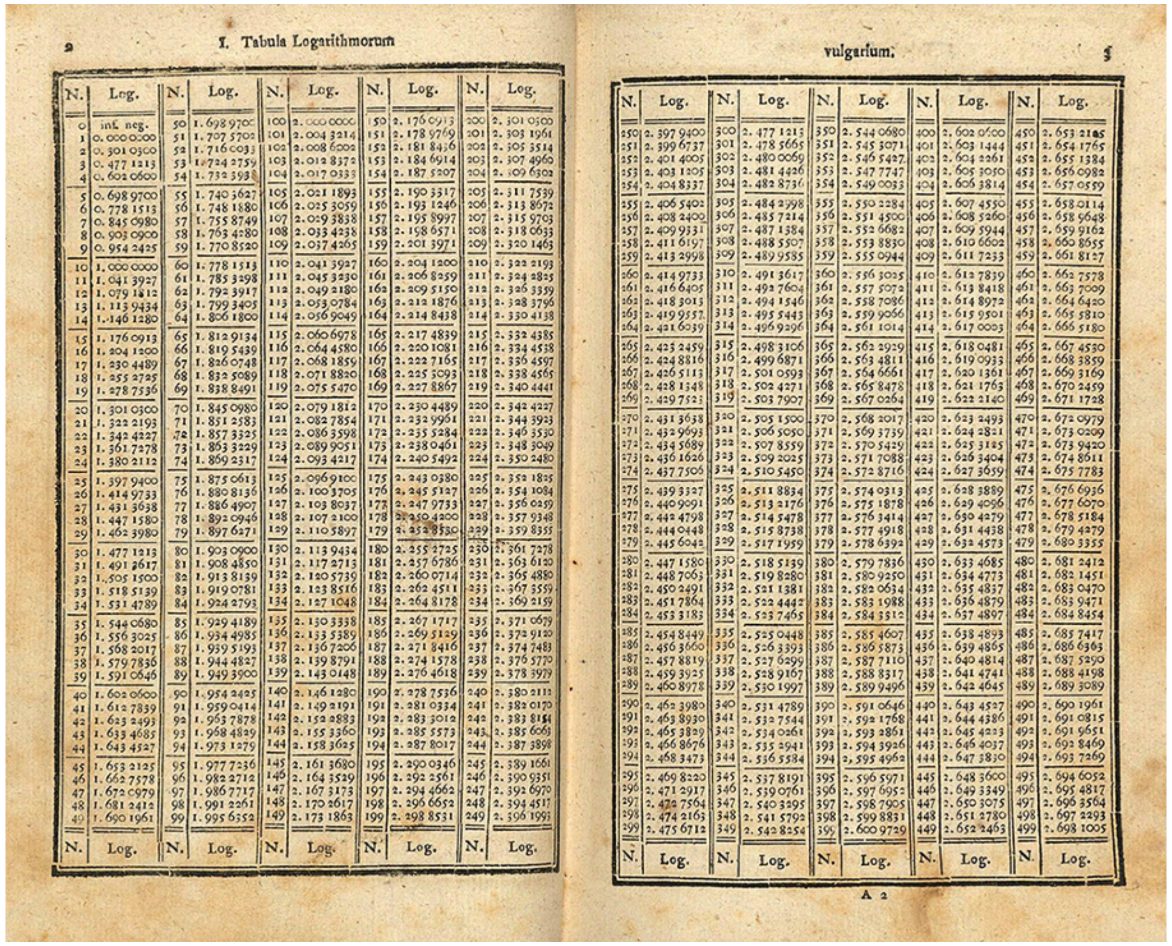


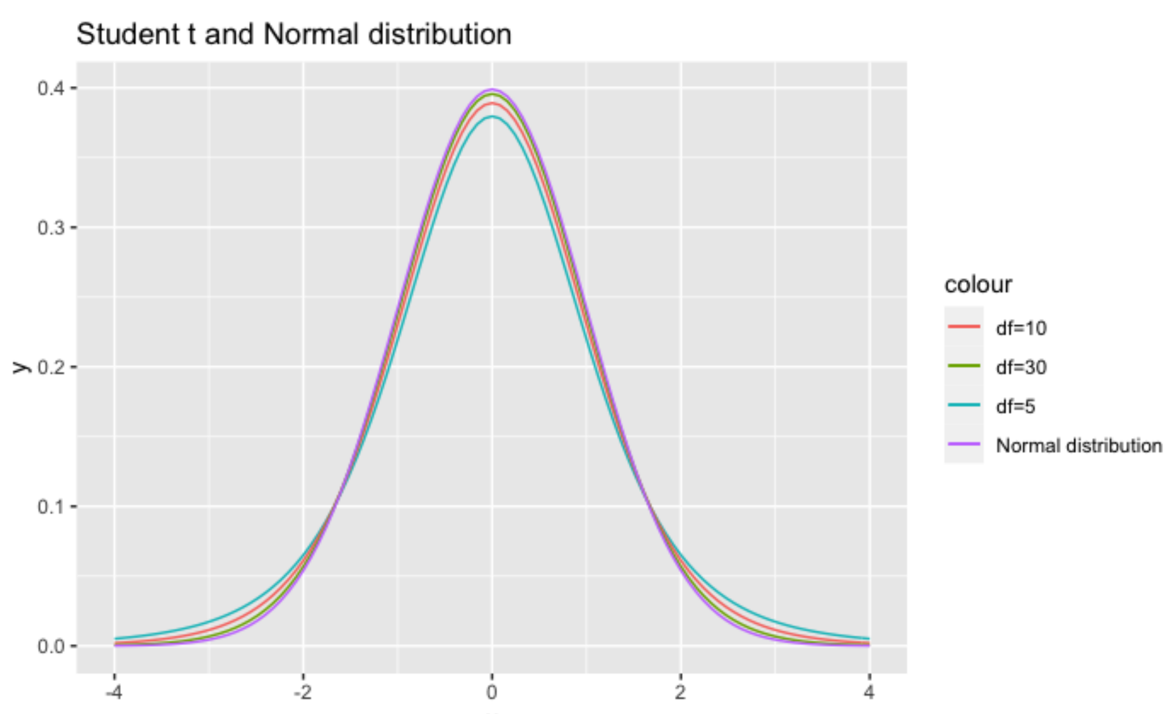



# Mathematical table



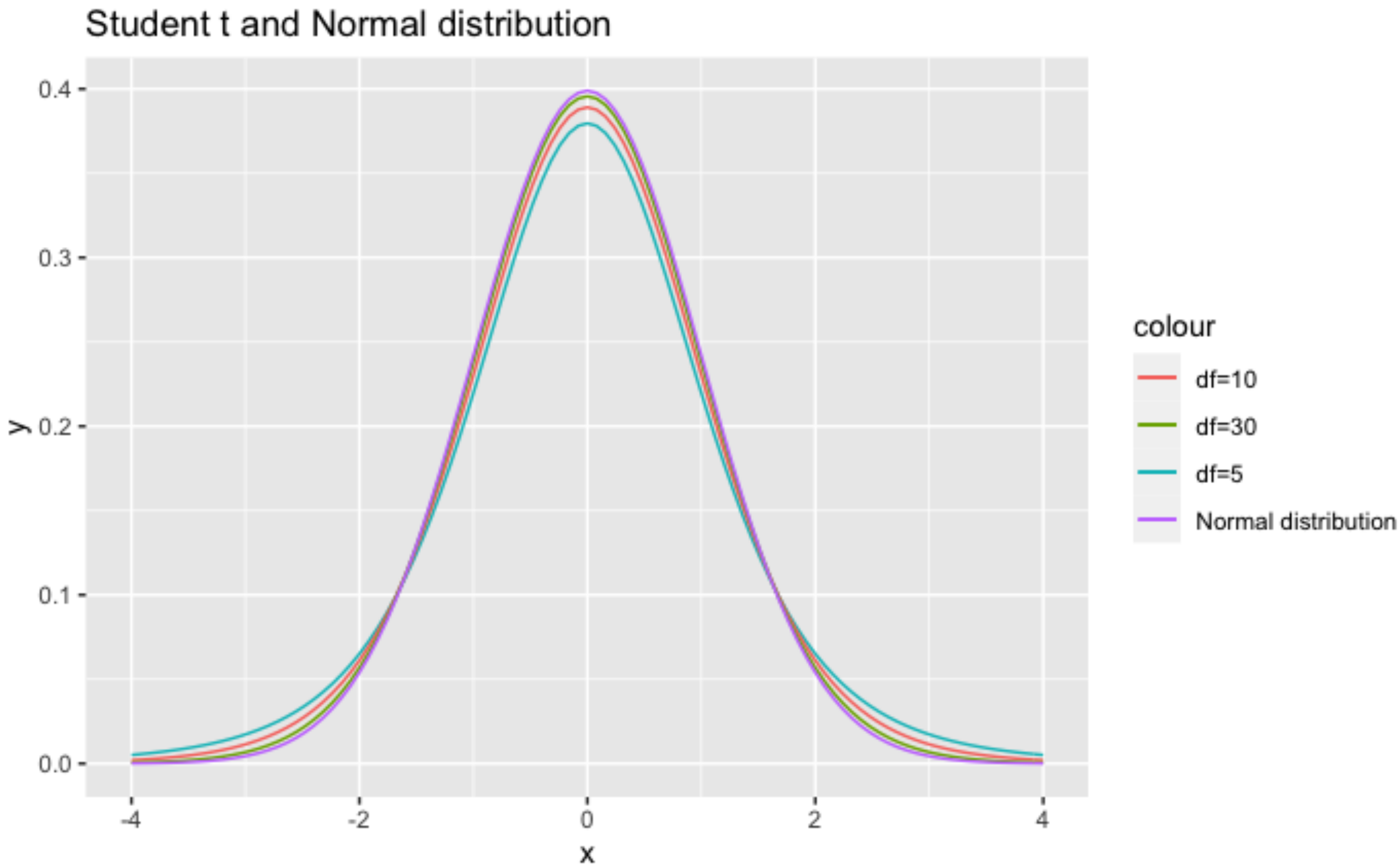




Z-value

| z | .00 |
|---|---|
| 0.0 | .5000 |
| 0.1 | .5398 |
| 0.2 | .5793 |
| 0.3 | .6179 |
| 0.4 | .6554 |
| 0.5 | .6915 |
| 0.6 | .7257 |
| 0.7 | .7580 |
| 0.8 | .7881 |
| 0.9 | .8159 |
| 1.0 | .8413 |
| 1.1 | .8643 |
| 1.2 | .8849 |
| 1.3 | .9032 |
| 1.4 | .9192 |
| 1.5 | .9332 |
| 1.6 | .9452 |
| 1.7 | .9554 |
| 1.8 | .9641 |
| 1.9 | .9713 |
| 2.0 | .9772 |
| 2.1 | .9821 |
| 2.2 | .9861 |
| 2.3 | .9893 |
| 2.4 | .9918 |
| 2.5 | .9938 |
| 2.6 | .9953 |
| 2.7 | .9965 |
| 2.8 | .9974 |
| 2.9 | .9981 |
| 3.0 | .9987 |
| 3.1 | .9990 |
| 3.2 | .9993 |
| 3.3 | .9995 |
| 3.4 | .9997 |

T-value

| *One-sided* | 75% | 80% | 85% | 90% | 95% | 97.5% | 99% | 99.5% | 99.75% | 99.9% | 99.95% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Two-sided* | 50% | 60% | 70% | 80% | 90% | 95% | 98% | 99% | 99.5% | 99.8% | 99.9% |
| 1 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 127.3 | 318.3 | 636.6 |
| 2 | 0.816 | 1.080 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 14.09 | 22.33 | 31.60 |
| 3 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 7.453 | 10.21 | 12.92 |
| 4 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| 17 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| 18 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.197 | 3.610 | 3.922 |
| 19 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| 20 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |
| 21 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 |
| 22 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.119 | 3.505 | 3.792 |
| 23 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.104 | 3.485 | 3.767 |
| 24 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.091 | 3.467 | 3.745 |
| 25 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.078 | 3.450 | 3.725 |
| 26 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.067 | 3.435 | 3.707 |
| 27 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.057 | 3.421 | 3.690 |
| 28 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.047 | 3.408 | 3.674 |
| 29 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.038 | 3.396 | 3.659 |
| 30 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.030 | 3.385 | 3.646 |
| 40 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 2.971 | 3.307 | 3.551 |
| 50 | 0.679 | 0.849 | 1.047 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 2.937 | 3.261 | 3.496 |
| 60 | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 2.915 | 3.232 | 3.460 |
| 80 | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 2.887 | 3.195 | 3.416 |
| 100 | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 2.871 | 3.174 | 3.390 |
| 120 | 0.677 | 0.845 | 1.041 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 2.860 | 3.160 | 3.373 |
| ∞ | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 2.807 | 3.090 | 3.291 |


Student t and Normal distribution
0.4
0.3
0.2
0.1
0.0
y
-4
-2
0
2
4
x
colour
df=10
df=30
df=5
Normal distribution

Normal (Gaussian) distribution

$$Z \sim N(\mu, \sigma^2)$$

Chi-square (df)

$$If\ Z \sim N(0,1), and\ X = \sum_i^k (Z_i^2), then\ X \sim \chi^2(k)$$

Student T(df)

$$If\ Z \sim N(0,1),\ U \sim \chi^2(n),\ and\ X = \frac{Z}{\sqrt{U/n}},\ then\ X \sim t(n)$$

F(df1, df2)

$$If\ U1 \sim \chi^2(n1),\ U2 \sim \chi^2(n2),\ and\ X = \frac{U1/n1}{U2/n2},\ then\ X \sim F(n1,n2)$$

$$t(n)^2 = (\frac{Z}{\sqrt{U/n}})^2 = \frac{Z^2}{U/n} = F(1,n)$$

# Statistics

## Two concepts and Two Theorems

Sample & Population

Theorem 1: Law of Large Numbers

Theorem 2: Central Limit Theorem

## Important Distributions

Normal (mu, sigma2)

Chi-square (df)

T(df)

F(df1, df2)

## Hypothesis testing and Two types of errors

Null hypothesis (H0) & Alternative hypothesis (H1)

False Positive & False Negative Errors

## Statistical tests

One-sample t test

Two-sample t-test

Paired t-test

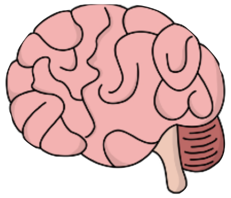
ANOVA

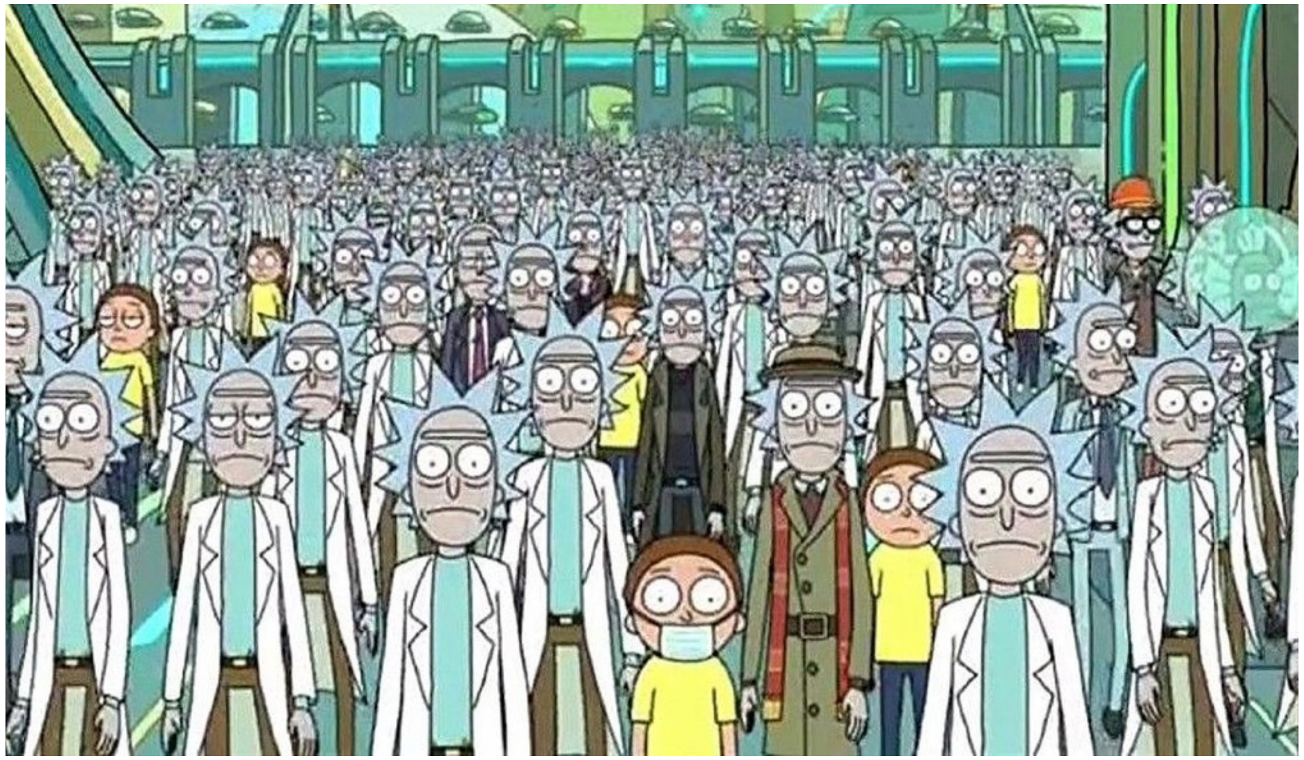
Repeated measure ANOVA

Regression

**Sample:** A fraction

**Population:** A whole

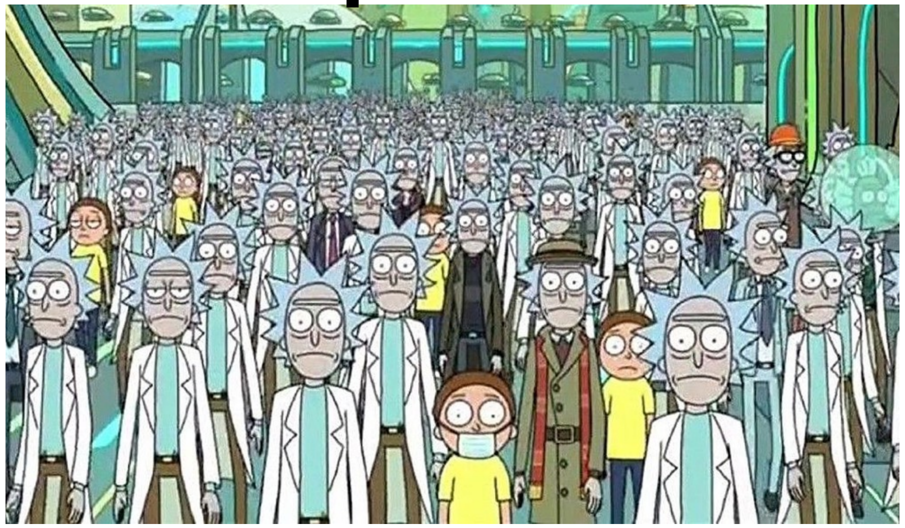




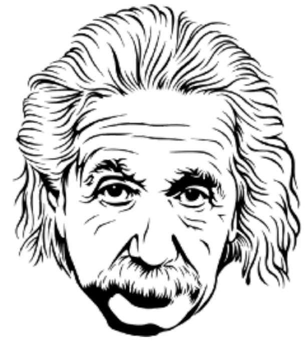
# One-sample T test

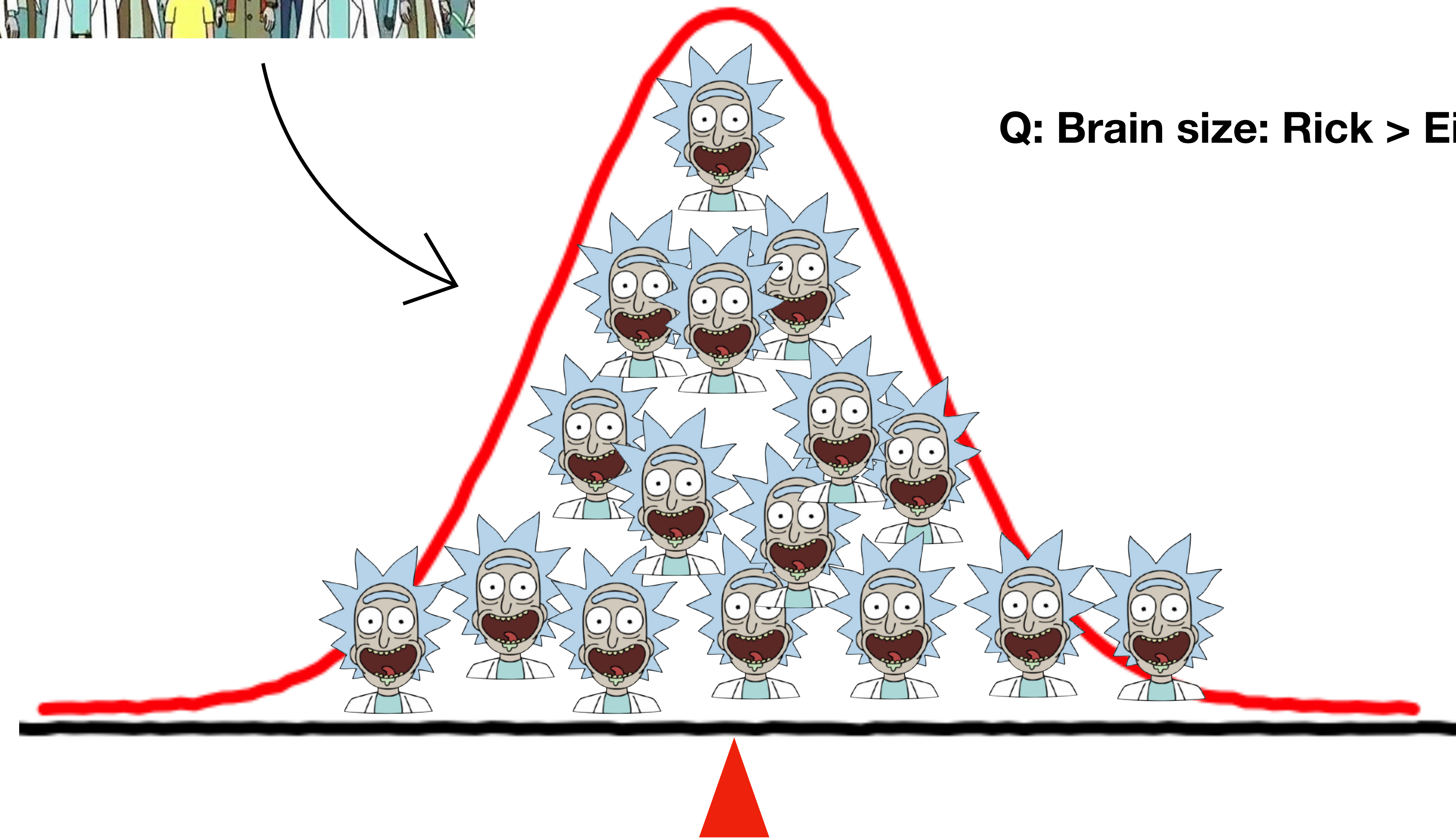
## Population

## One Sample T-test



?





**Q: Brain size: Rick > Einstein?**

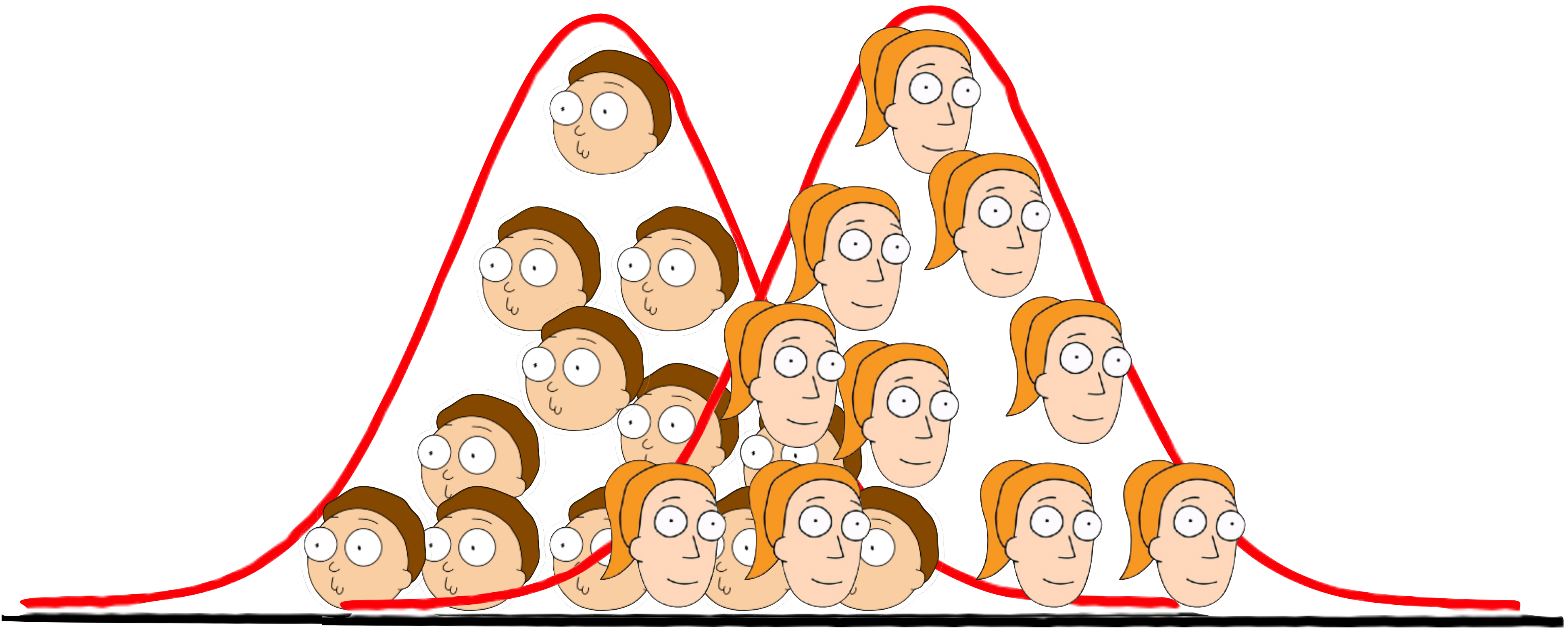
H0 (Null Hypothesis): Rick=Einstein

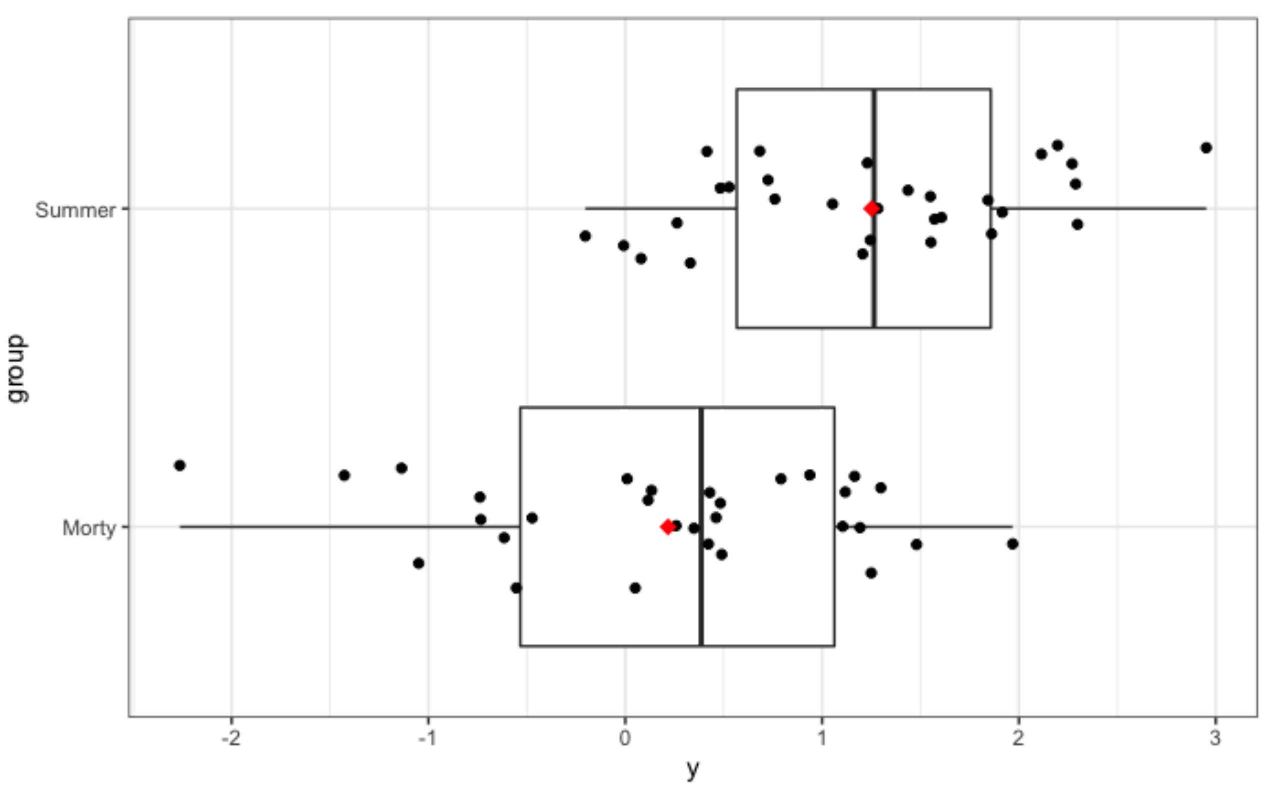
H1 (Alternative Hypothesis): Rick!=Einstein

Two-sample T test
One way ANOVA
Linear regression

Q: Difference: Summer > Morty?

# Two Sample T-test

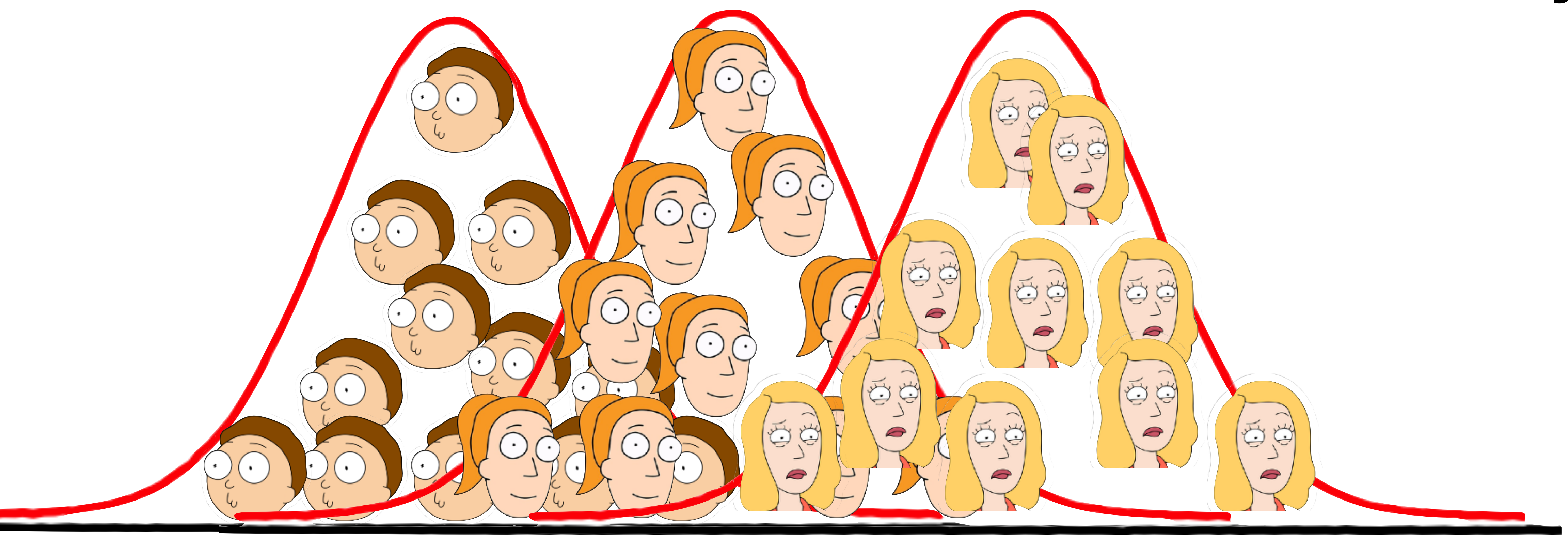


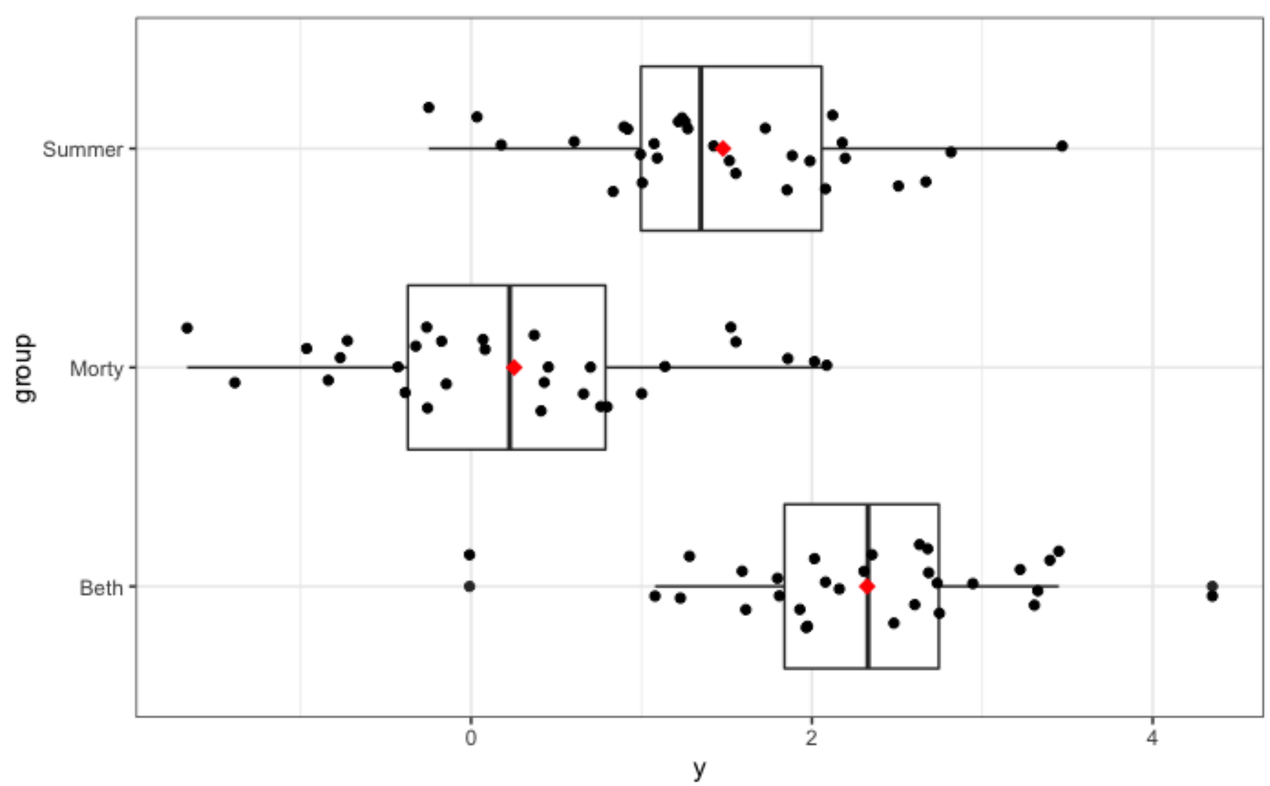



Two Sample t-test

data:  y1 and y2
t = -5.2098, df = 58, p-value = 2.616e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6962870 -0.7545972
sample estimates:
mean of x mean of y
0.2528962 1.4783383

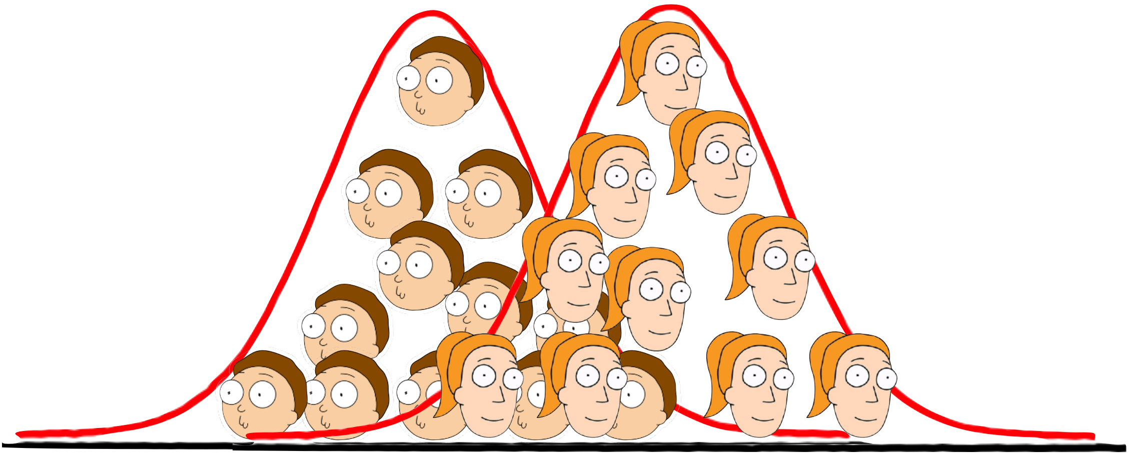
# One-way ANOVA





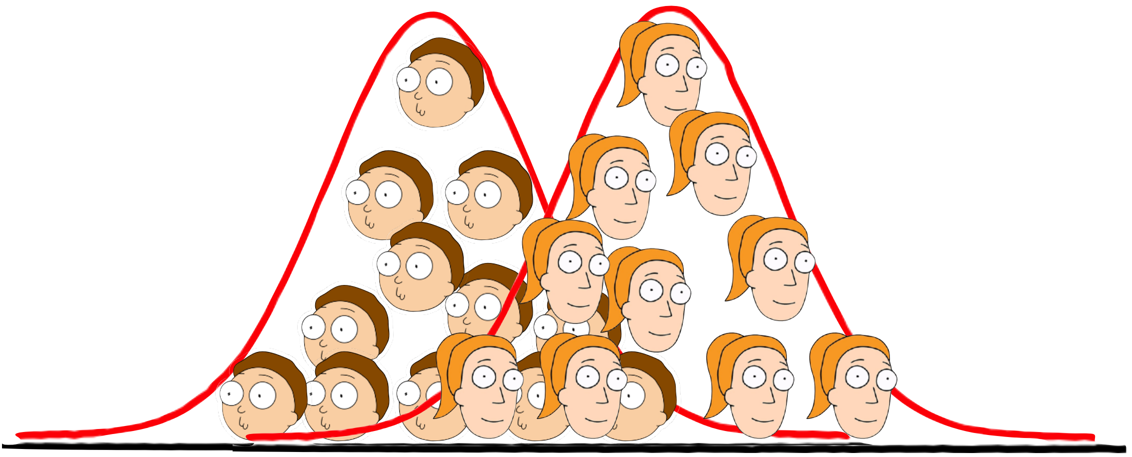

```
Anova Table (Type II tests)

Response: y
          Sum Sq Df F value    Pr(>F)
group     65.088  2  40.404 3.889e-13 ***
Residuals 70.076 87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

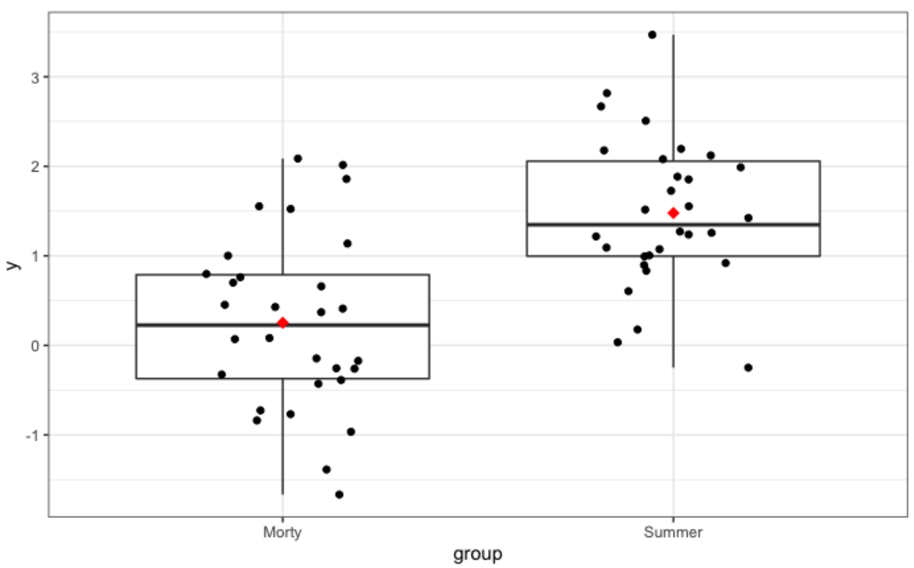
# Two Sample T-test  as One-way (2 levels) ANOVA

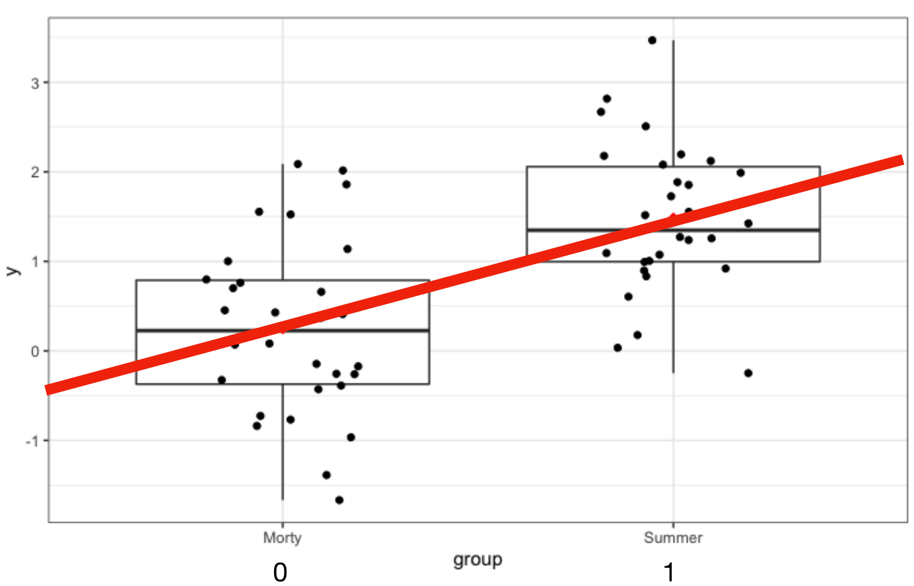


```
        Two Sample t-test

data:  y1 and y2
t = -5.2098, df = 58, p-value = 2.616e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6962870 -0.7545972
sample estimates:
mean of x mean of y
0.2528962 1.4783383

Anova Table (Type II tests)

Response: y
          Sum Sq Df F value    Pr(>F)
group     22.526  1  27.142 2.616e-06 ***
Residuals 48.136 58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# Two Sample T-test as linear regression






**Two Sample t-test**

```
data:  y1 and y2
t = -5.2098, df = 58, p-value = 2.616e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6962870 -0.7545972
sample estimates:
mean of x mean of y
0.2528962 1.4783383
```

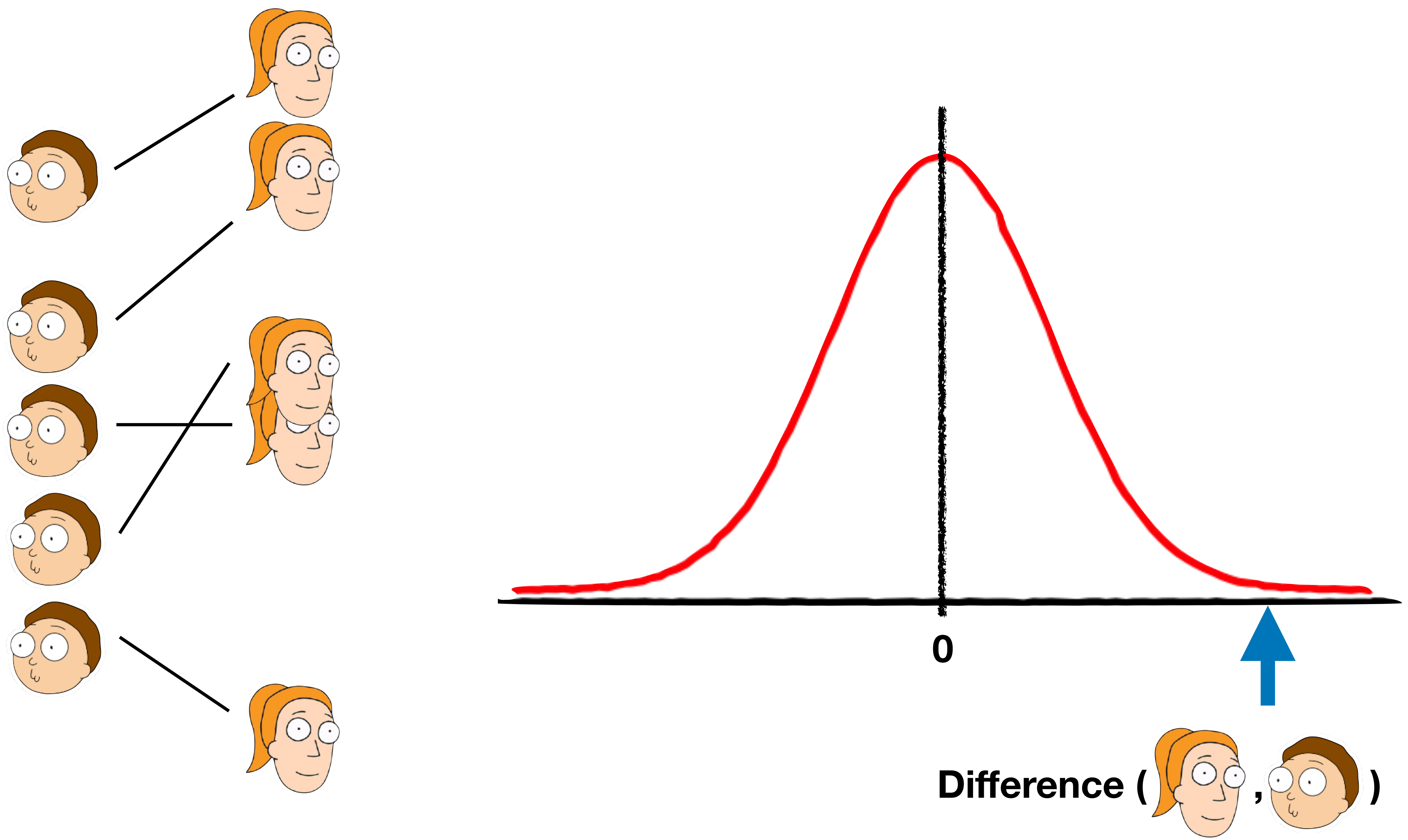
**Anova Table (Type II tests)**

```
Response: y
          Sum Sq Df F value    Pr(>F)
group     22.526  1  27.142 2.616e-06 ***
Residuals 48.136 58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = y ~ 1 + group, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9195 -0.5636 -0.1127  0.5591  1.9906

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2529     0.1663    1.52    0.134
groupSummer   1.2254     0.2352    5.21 2.62e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.911 on 58 degrees of freedom
Multiple R-squared:  0.3188,    Adjusted R-squared:  0.307
F-statistic: 27.14 on 1 and 58 DF,  p-value: 2.616e-06
```

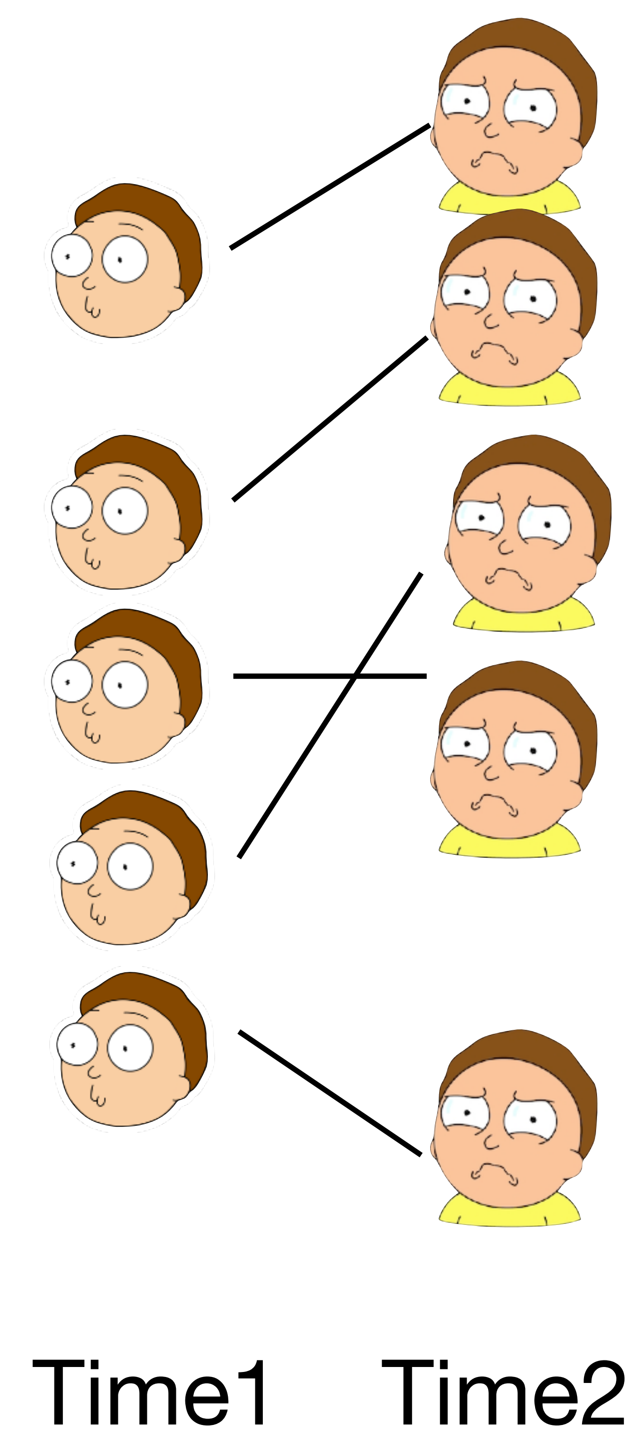
Paired t
Repeated Measure ANOVA
Linear Regression

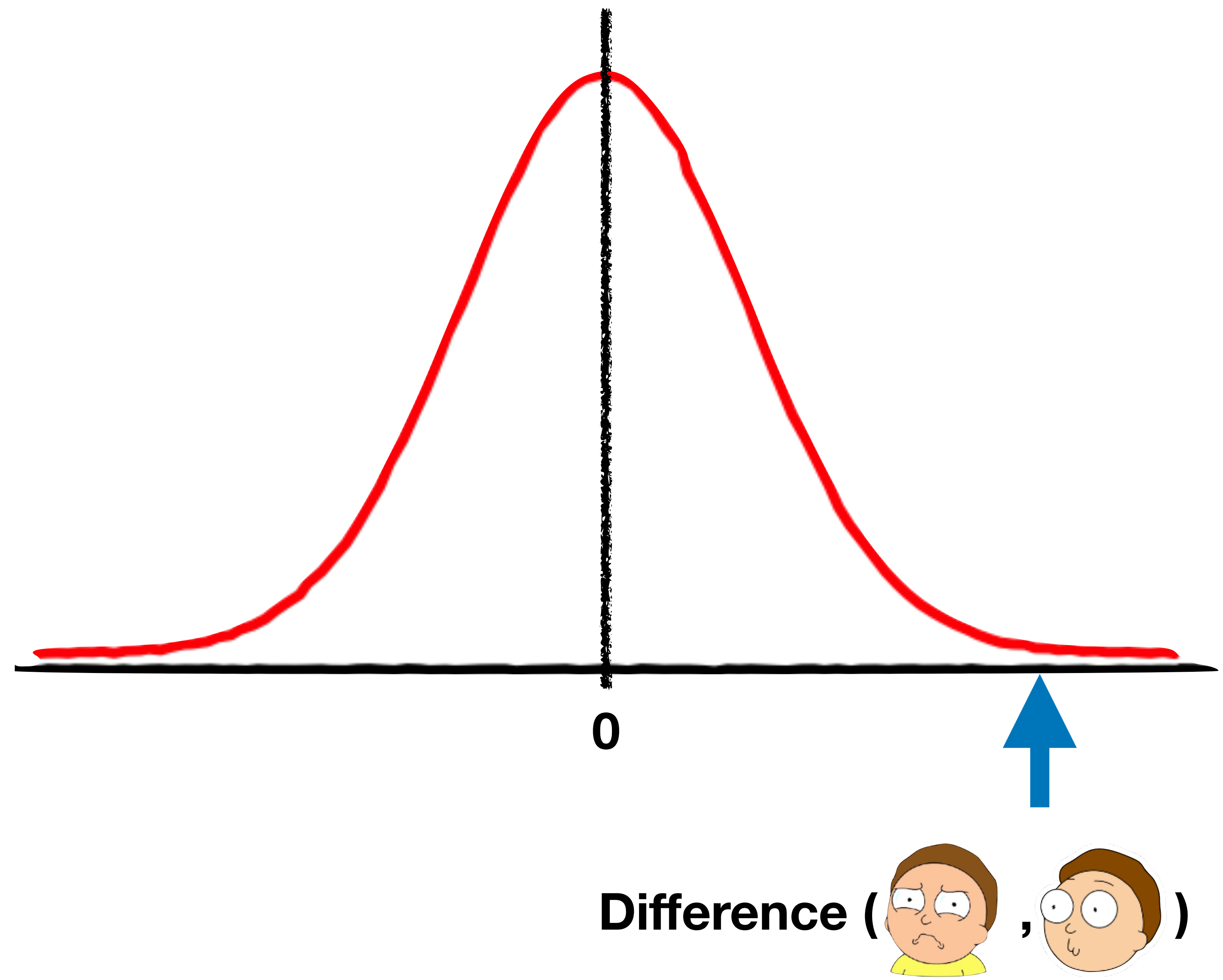
# Paired Sample T-test



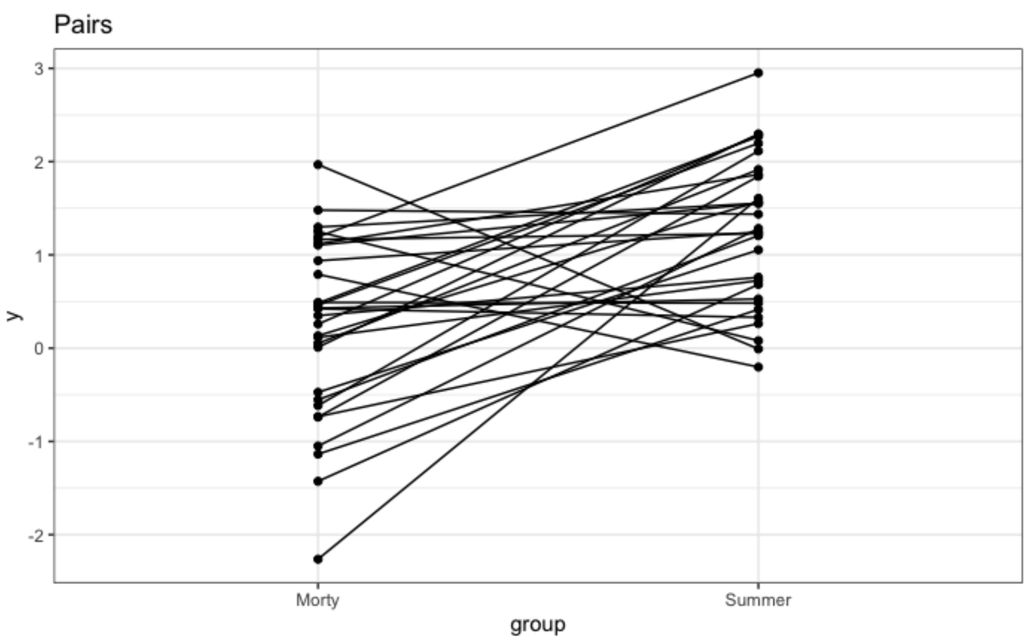
# Paired T-test

Repeated Measure



# Paired Sample T-test as One-sample T-test



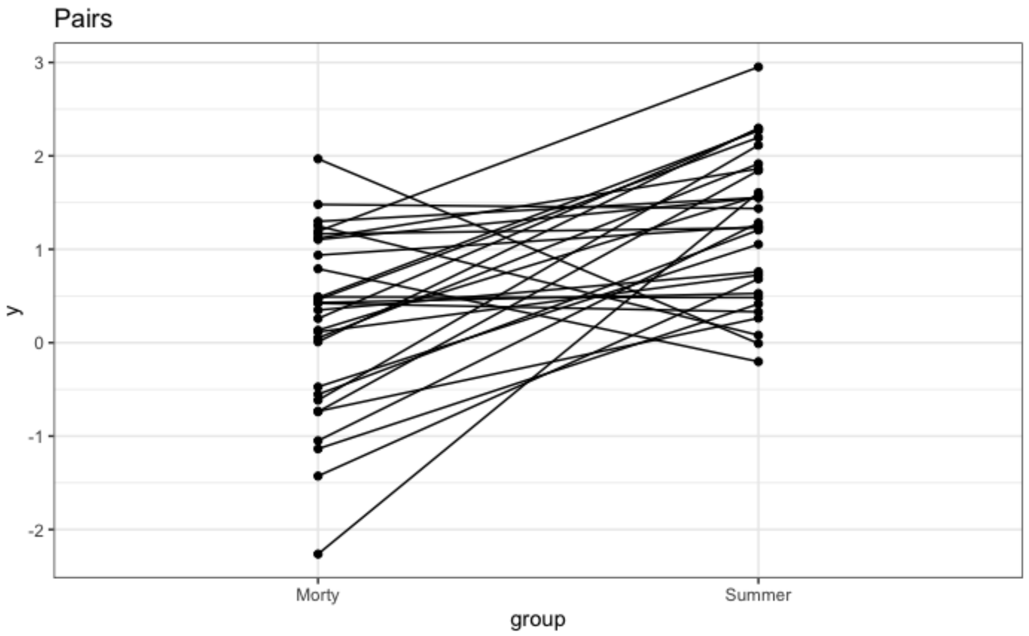

Paired t-test

data: y1 and y2
t = -4.8467, df = 29, p-value = 3.884e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.7425597 -0.7083245
sample estimates:
mean of the differences
-1.225442

One Sample t-test

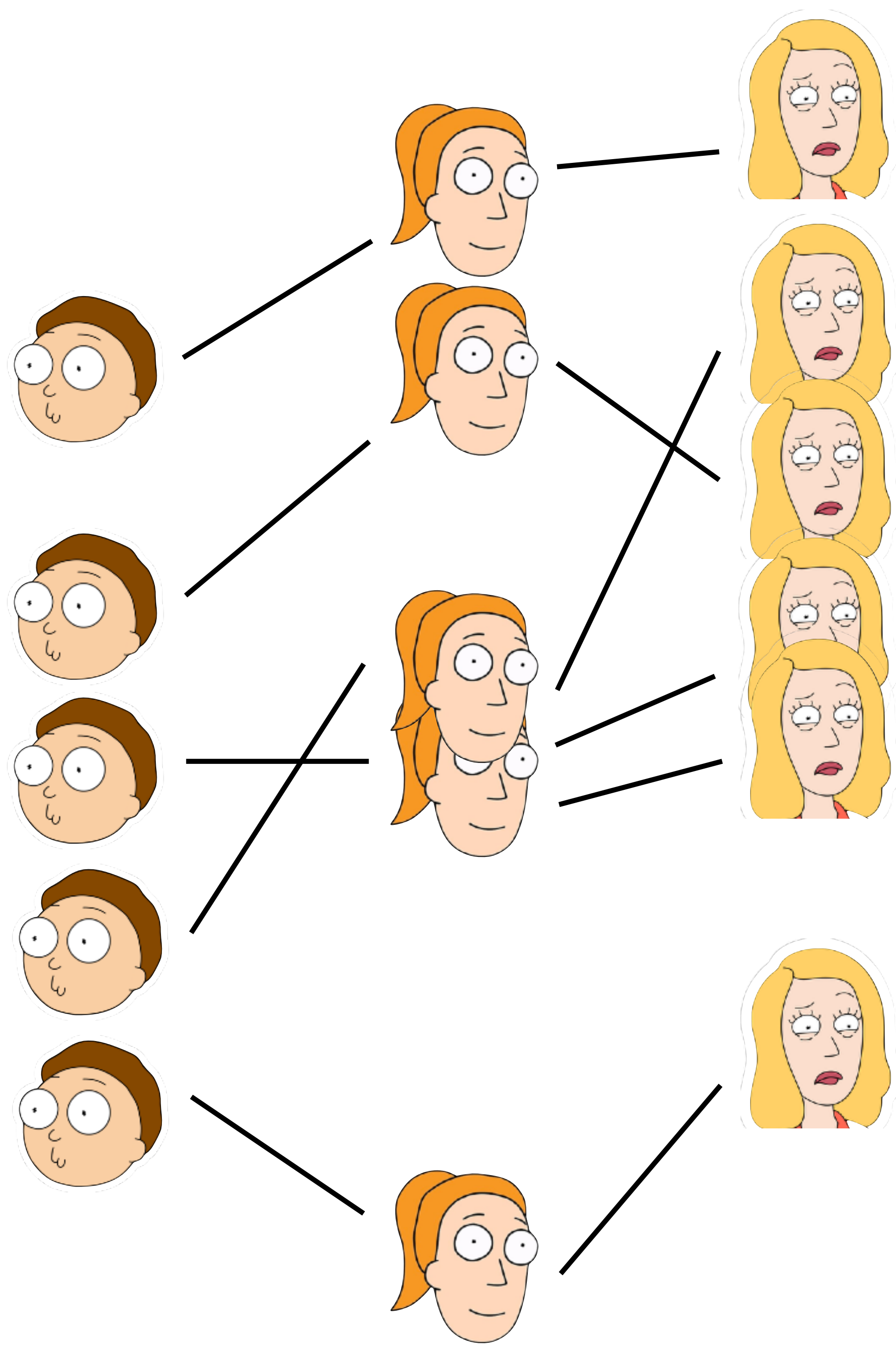
data: y1 - y2
t = -4.8467, df = 29, p-value = 3.884e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-1.7425597 -0.7083245
sample estimates:
mean of x
-1.225442

# Paired Sample T-test as repeated measure ANOVA

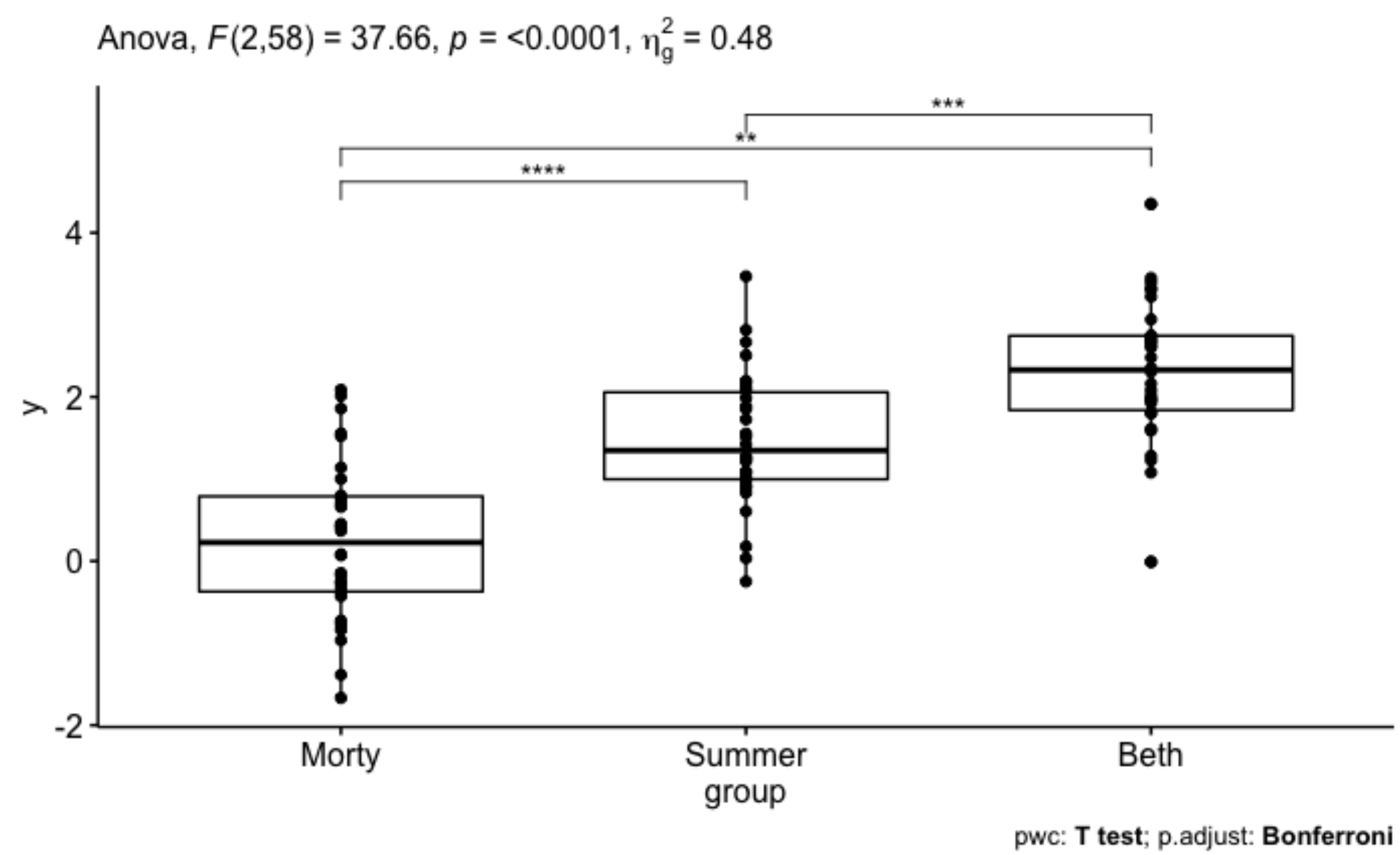



```
Paired t-test

data:  y1 and y2
t = -4.8467, df = 29, p-value = 3.884e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7425597 -0.7083245
sample estimates:
mean of the differences
              -1.225442

ANOVA Table (type III tests)

  Effect DFn DFd     F        p p<.05   ges
1  group   1  29 23.49 3.88e-05     * 0.319
```

# One-way repeated measure ANOVA



```
ANOVA Table (type III tests)

  Effect DFn DFd      F        p p<.05   ges
1  group   2  58 37.664 3.29e-11     * 0.482
```

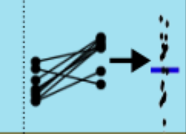
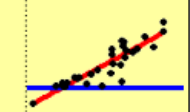
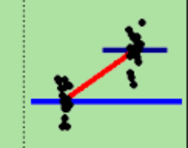
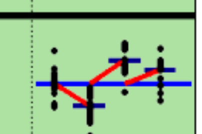
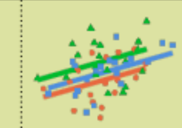


# https://lindeloev.github.io/tests-as-linear/

# Common statistical tests are linear models

*Last updated: 02 April, 2019*

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

| | Common name | Built-in function in R | Equivalent linear model in R | Exact? | The linear model in words | Icon |
|---|---|---|---|---|---|---|
| **Simple regression: lm(y ~ 1 + x)** | **y is independent of x**<br>P: One-sample t-test<br>N: Wilcoxon signed-rank | t.test(y)<br>wilcox.test(y) | lm(y ~ 1)<br>lm(signed_rank(y) ~ 1) | ✓<br>for N >14 | One number (intercept, i.e., the mean) predicts **y**.<br>- (Same, but it predicts the *signed rank* of **y**.) |  |
| | P: Paired-sample t-test<br>N: Wilcoxon matched pairs | t.test($y_1$, $y_2$, paired=TRUE)<br>wilcox.test($y_1$, $y_2$, paired=TRUE) | lm($y_2$ - $y_1$ ~ 1)<br>lm(signed_rank($y_2$ - $y_1$) ~ 1) | ✓<br>for N >14 | One intercept predicts the pairwise **$y_2$-$y_1$** differences.<br>- (Same, but it predicts the *signed rank* of **$y_2$-$y_1$**.) |  |
| | **y ~ continuous x**<br>P: Pearson correlation<br>N: Spearman correlation | cor.test(x, y, method='Pearson')<br>cor.test(x, y, method='Spearman') | lm(y ~ 1 + x)<br>lm(rank(y) ~ 1 + rank(x)) | ✓<br>for N >10 | One intercept plus **x** multiplied by a number (slope) predicts **y**.<br>- (Same, but with *ranked* **x** and **y**) |  |
| | **y ~ discrete x**<br>P: Two-sample t-test<br>P: Welch's t-test<br>N: Mann-Whitney U | t.test($y_1$, $y_2$, var.equal=TRUE)<br>t.test($y_1$, $y_2$, var.equal=FALSE)<br>wilcox.test($y_1$, $y_2$) | lm(y ~ 1 + $G_2$)[A]<br>gls(y ~ 1 + $G_2$, weights=...[B])[A]<br>lm(signed_rank(y) ~ 1 + $G_2$)[A] | ✓<br>✓<br>for N >11 | An intercept for **group 1** (plus a difference if **group 2**) predicts **y**.<br>- (Same, but with one variance *per group* instead of one common.)<br>- (Same, but it predicts the *signed rank* of **y**.) |  |
| **Multiple regression: lm(y ~ 1 + $x_1$ + $x_2$ +...)** | P: One-way ANOVA<br>N: Kruskal-Wallis | aov(y ~ group)<br>kruskal.test(y ~ group) | lm(y ~ 1 + $G_2$ + $G_3$ +...+ $G_N$)[A]<br>lm(rank(y) ~ 1 + $G_2$ + $G_3$ +...+ $G_N$)[A] | ✓<br>for N >11 | An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**.<br>- (Same, but it predicts the *rank* of **y**.) |  |
| | P: One-way ANCOVA | aov(y ~ group + x) | lm(y ~ 1 + $G_2$ + $G_3$ +...+ $G_N$ + x)[A] | ✓ | - (Same, but plus a slope on **x**.)<br>*Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.* |  |
| | P: Two-way ANOVA | aov(y ~ group * sex) | lm(y ~ 1 + $G_2$ + $G_3$ + ... + $G_N$ +<br>$S_2$ + $S_3$+ ... + $S_K$ +<br>$G_2$*$S_2$+$G_3$*$S_3$+...+$G_N$*$S_K$) | ✓ | Interaction term: changing **sex** changes the **y ~ group** parameters.<br>*Note: **$G_{2 \text{ to } N}$** is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for **$S_{2 \text{ to } K}$** for sex. The first line (with $G_i$) is main effect of group, the second (with $S_i$) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "$S_2$" and line 3 would be $S_2$ multiplied with each $G_i$.* | *[Coming]* |
| | **Counts ~ discrete x**<br>N: Chi-square test | chisq.test(groupXsex_table) | **Equivalent log-linear model**<br>glm(y ~ 1 + $G_2$ + $G_3$ + ... + $G_N$ +<br>$S_2$ + $S_3$+ ... + $S_K$ +<br>$G_2$*$S_2$+$G_3$*$S_3$+...+$G_N$*$S_K$, family=...)[A] | ✓ | Interaction term: (Same as Two-way ANOVA.)<br>*Note: Run glm using the following arguments:* `glm(model, family=poisson())`<br>*As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where $\alpha_i$ and $\beta_j$ are proportions. See more info in the accompanying notebook.* | *Same as Two-way ANOVA* |
| | N: Goodness of fit | chisq.test(y) | glm(y ~ 1 + $G_2$ + $G_3$ +...+ $G_N$, family=...)[A] | ✓ | (Same as One-way ANOVA and see Chi-Square note.) | *1W-ANOVA* |

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for y = 1·b + a·x which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables $G_i$ and $S_i$ are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when Δx = 1 between categories the difference equals the slope. Subscripts (e.g., $G_2$ or $y_1$) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G₂, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net