# R NOTES

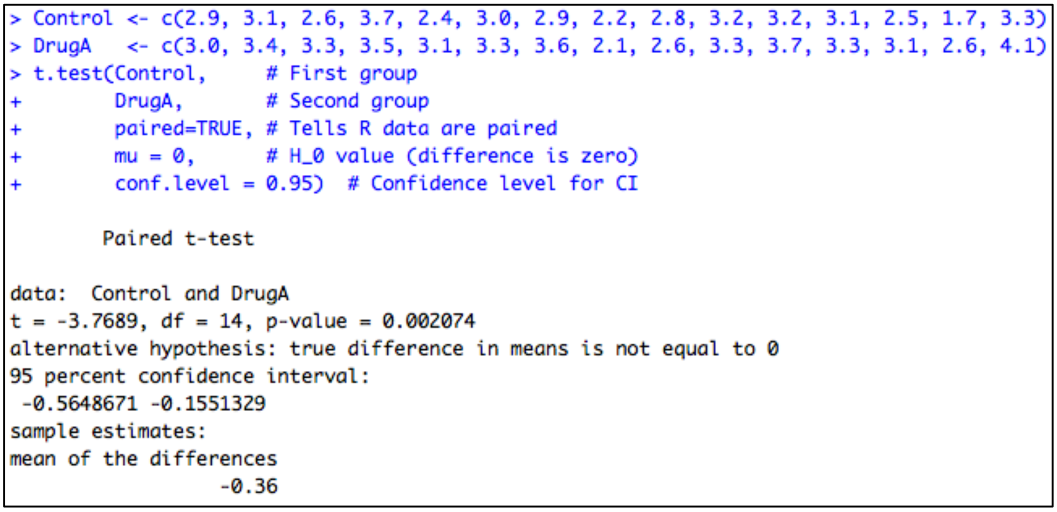
**Section 1**: Summaries, ITQ, plotting, distributions and descriptions (location, spread)
**Section 2**: Confidence Intervals, different distributions (normal, t, binom, poisson), estimations
**Section 3**: Hypothesis testing, comparing two groups, tests (z, t, …)

## SECTION 1

| | |
|---|---|
| **Manually Entering Data**<br>`X <- c(1,2,3,4,5,6,7,8)`<br>`Test <- data.frame(x,y)` | ‘c’ stands for combine and stores the values as a column called ‘x’<br>Creates a data set called ‘Test’ with two columns ‘x’ and ‘y’ |
| **Basic Functions**<br>`Diff <- x - y`<br>`Test <- data.frame(Test, Diff)` | Creates a set that is the difference between ‘x’ and ‘y’ and adds a column to the whole dataset |
| `Sqrt(x)`<br>`Log(x)`<br>`Sum(x)`<br>`Mean(x)`<br>$\texttt{Mean} = \bar{x} = \text{sample mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = 56/10;$<br>`Median(x)`<br>$\texttt{Median} = \hat{c}_{0.5} = \text{sample median} = \hat{c}_{0.5} = 5;$<br>`Min(x)`<br>`Sd(x)`<br>$\texttt{StDev} = s$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2;$<br>(sample standard deviation) [ $(\bar{x}-2s, \bar{x}+2s)$ contains about 95% of the sample.] | Natural log used |
| `Attach(x)` | Used to make things easier, instead of using dataset$column, can just refer straight to column |
| **5-number summary of a column of variables**<br>`Summary(data)`<br>`> summary(x)`<br>`   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.`<br>`   1.00    4.00    5.00    5.60    6.75   12.00` | Provides a 5 number summary of data |
| **Interquartile Range**<br>`IQR(x)`<br>**sample interquartile range:** $\hat{\tau}$ = IQR = Q3 – Q1 or $\hat{c}_{0.75} - \hat{c}_{0.25}$.<br>It is a single number: it is the difference, and not the interval. | Uses the function for output<br>Uses a different function that can be adapted for different quantile levels |

## Comparing paired and independent samples

`t.test(x,y, paired=TRUE, mu=?, conf.level=?)`



```
> Control <- c(2.9, 3.1, 2.6, 3.7, 2.4, 3.0, 2.9, 2.2, 2.8, 3.2, 3.2, 3.1, 2.5, 1.7, 3.3)
> DrugA   <- c(3.0, 3.4, 3.3, 3.5, 3.1, 3.3, 3.6, 2.1, 2.6, 3.3, 3.7, 3.3, 3.1, 2.6, 4.1)
> t.test(Control,     # First group
+        DrugA,       # Second group
+        paired=TRUE, # Tells R data are paired
+        mu = 0,      # H_0 value (difference is zero)
+        conf.level = 0.95)  # Confidence level for CI

        Paired t-test

data:  Control and DrugA
t = -3.7689, df = 14, p-value = 0.002074
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5648671 -0.1551329
sample estimates:
mean of the differences 
                  -0.36 
```

where paired=TRUE indicates the two groups are matched, so that they're independent and for comparison, mu = ? is the $H_0$ value (difference = 0)

## How to ANSWER HYPOTHESIS TESTING QUESTIONS

### Hypotheses:

$H_0$: - true mean ($\mu$) is equal to 1.3

- Mean of control = mean of drug
- Mean of sample 1 = mean of sample 2

$H_{a/1}$: - the same but *not equal* instead

### Assuming $H_0$ is true, state the distribution of the estimator ($\bar{X}$):

- Normal distribution (CLT)
- Binomial distribution (proportions)
- Poisson distribution (rates)

**Test statistic:** Any value that R outputs

**p-value:** Value that determines accept/do not accept

### Conclusion in the context of these data:

MOST IMPORTANT

- We do not accept/do not accept $H_0$ as p-value<0.05/p-value>0,05
- The aim was to determine if there was any significant association between *x* and *y*: Yes, men with *x* corresponded with lower *y* measurements.
- Therefore, the group/sample had lower/higher/different mean value than the general population/null hypothesis and is significant.

**Test statistic**: a standardized value that is calculated from sample data during a hypothesis test. You can use test statistics to determine whether to reject the null hypothesis. The test statistic compares your data with what is expected under the null hypothesis.

**p-value:** This probability represents the likelihood of obtaining a sample mean that is at least as extreme as our sample mean in both tails of the