# A Brief Review of Univariate Statistics

Today's goals:

- Inferential statistics as signal detection
- Two groups
- More than two groups

The primary goal of inferential statistics is to provide a basis for confidently claiming that a theory-predicted "signal" is present in the data.

Statistical tests can be thought of as "signal-to-noise" ratios:

$$Test = \frac{Signal + Noise}{Noise}$$

When the signal is "large enough" relative to the noise, we may claim it to be a "significant" signal.

Another way to frame the basic inferential test is as an "effect" relative to its standard error.

$$Test = \frac{Signal + Noise}{Noise} = \frac{Effect}{SE_{Effect}}$$

The standard error (SE), or standard deviation of the effect, tells us how much the effect can be expected to vary under the null hypothesis that there is no signal. If the effect is unusual in this context, we can claim there is probably a signal present (i.e., not just noise).

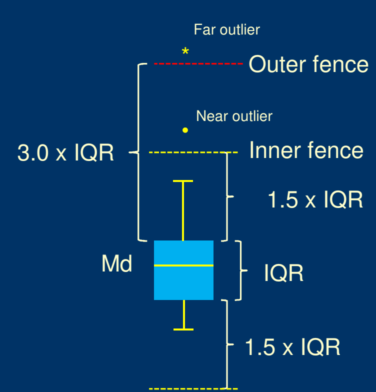
A two-group design and the t-test provide a simple example.

The effect or signal of interest is the mean difference:

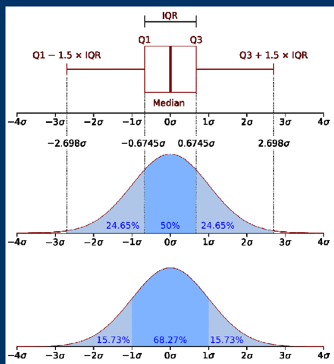
$6.3 – 4.7 = 1.6$

Is this big or "significant?" We need to consider variability under the null to know this.

| | Treatment | Control |
|---|---|---|
| Subject 1 | 8 | 7 |
| Subject 2 | 7 | 5 |
| Subject 3 | 5 | 4 |
| Subject 4 | 6 | 5 |
| Subject 5 | 5 | 3 |
| Subject 6 | 7 | 5 |
| Subject 7 | 6 | 4 |
| Subject 8 | 7 | 3 |
| Subject 9 | 4 | 6 |
| Subject 10 | 8 | 5 |
| | | |
| Mean | 6.3 | 4.7 |
| SE | .42 | .40 |

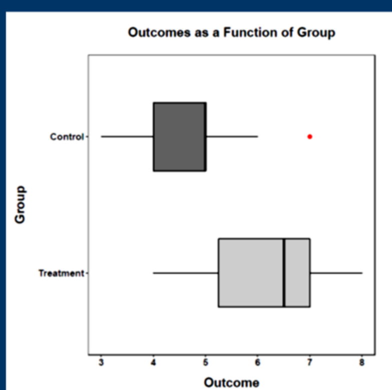



The boxplot provides very basic information about group separation relative to group variability. It also displays distribution shape and presence of unusual cases.




The boxplot can be compared to the normal distribution to give some sense of expected values, especially outliers.

https://en.wikipedia.org/wiki/Box_plot




The two-group problem can be cast in the language of matrix algebra. If **X** is a vector of group means:

$$X = \begin{bmatrix} 6.3 \\ 4.7 \end{bmatrix}$$

and **L** is a vector containing the weights for the linear combination of means:

$$L = \begin{bmatrix} 1 & -1 \end{bmatrix}$$

Then **LX** gives the mean difference:

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 6.3 \\ 4.7 \end{bmatrix} = 1.6$$

The variance-covariance matrix for the means (Σ) is:

$$\sum = \begin{bmatrix} .18 & .00 \\ .00 & .16 \end{bmatrix}$$

Σ is a diagonal matrix because the groups are assumed to be independent

The variance of a linear combination is given by:

$$\sigma_{LC}^2 = \begin{bmatrix} W_1 & W_2 & \dots & W_k \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,k} \\ \sigma_{2,1} & \sigma_{2,2}^2 & \dots & \sigma_{2,k} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{k,1} & \sigma_{k,2} & \dots & \sigma_{k,k}^2 \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \\ \cdot \\ \cdot \\ \cdot \\ W_k \end{bmatrix}$$

so, the variance of the mean difference is:

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} .18 & .00 \\ .00 & .16 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = .34$$

The ratio of the linear combination to its standard error is the t-test:

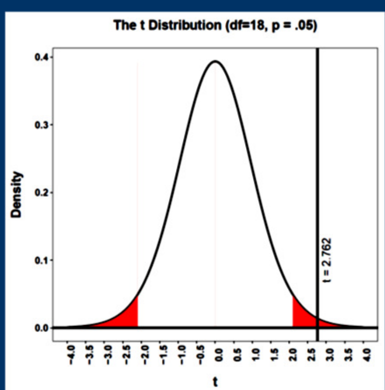
$$t = \frac{1.6}{\sqrt{.34}} = 2.76$$

```
 Two Sample t-test

data:  DV by Group
t = 2.8, df = 18, p-value = 0.01
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.383 2.817
sample estimates:
mean in group 1 mean in group 2
            6.3             4.7
```

The mean difference is large enough that we reject the null hypothesis, but what does that mean?

We use the concept of the *sampling distribution* to help us determine when an effect or signal is large enough to be more than just a manifestation of background noise or random variability.

Theoretical sampling distributions tell us what to expect if the null hypothesis is true, <u>and if assumptions about the data are true</u>.




In the context of the sampling distribution of mean differences and assuming the null hypothesis, the obtained mean difference is unusual.

The general equation:

$$t = \frac{LXM}{SE_{LXM}}$$

provides complete flexibility in deriving tests for effects that match those predicted by theory.

This focused approach is the difference between asking, “did anything happen?” and asking, “did this specific, predicted thing happen?”

This is especially important when there are more than two groups and more than one measure.

In the analysis of variance, the test of inference asks . . .

| $Group_1$ | $Group_2$ | $Group_3$ | $Group_4$ | $Group_5$ |
|---|---|---|---|---|
| $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |

is there more variability in these means than would be expected if the null hypothesis were true?

The null hypothesis is that the means are equal in the population and will vary in any sample only due to random variability.

One way of estimating how much variability to expect in the means, if the null hypothesis is true, is to use the variability within each condition (where the variability is assumed to be random).

The variance of a mean is related to the variance of the scores on which the mean is based:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{\sqrt{N}}$$

The F test in the analysis of variance examines the ratio of two different estimates of that random variability:

$$F \approx \frac{n\sigma^2_{\bar{X}}}{\overline{\sigma^2_X}}$$

If the null hypothesis is true, then the variance of the means should, when properly scaled, give an estimate of the variance of scores. The variance of scores can be gotten directly by pooling the separate within-group variances.

$$F \approx \frac{n\sigma_{\bar{X}}^2}{\overline{\sigma_X^2}}$$

$\hat{\sigma}_{\bar{X}}^2 \leftarrow$

| $Group_1$ | $Group_2$ | $Group_3$ | $Group_4$ | $Group_5$ |
|---|---|---|---|---|
| $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |
| $\hat{\sigma}_1^2$ | $\hat{\sigma}_2^2$ | $\hat{\sigma}_3^2$ | $\hat{\sigma}_4^2$ | $\hat{\sigma}_5^2$ |

$$F \approx \frac{n\sigma_{\bar{X}}^2}{\overline{\sigma_X^2}}$$

If the null hypothesis is true, this ratio will be close to 1.00.

If this ratio exceeds 1.00 “enough” (based on a sampling distribution), we reject the null hypothesis and assume there must be an additional source of variance in the means (the signal) that does not exist in the scores within groups. That signal is the between-group effect of the treatment.

The variance estimates used in the F ratio are obtained by *partitionin*g the total variability in the scores.

| $Group_1$ | $Group_2$ | $Group_3$ | $Group_4$ | $Group_5$ |
|---|---|---|---|---|
| $Y_{11}$ | $Y_{12}$ | $Y_{13}$ | $Y_{14}$ | $Y_{15}$ |
| $Y_{21}$ | $Y_{22}$ | $Y_{23}$ | $Y_{24}$ | $Y_{25}$ |
| $Y_{31}$ | $Y_{32}$ | $Y_{33}$ | $Y_{34}$ | $Y_{35}$ |
| $Y_{41}$ | $Y_{42}$ | $Y_{43}$ | $Y_{44}$ | $Y_{45}$ |
| $Y_{51}$ | $Y_{52}$ | $Y_{53}$ | $Y_{54}$ | $Y_{55}$ |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| $\bar{Y}_{.1}$ | $\bar{Y}_{.2}$ | $\bar{Y}_{.3}$ | $\bar{Y}_{.4}$ | $\bar{Y}_{.5}$ |

$$Y_{ij} - \overline{Y}_{..} = (Y_{ij} - \overline{Y}_{.j}) + (\overline{Y}_{.j} - \overline{Y}_{..})$$

| $Group_1$ | $Group_2$ | $Group_3$ | $Group_4$ | $Group_5$ |
|---|---|---|---|---|
| $Y_{11}$ | $Y_{12}$ | $Y_{13}$ | $Y_{14}$ | $Y_{15}$ |
| $Y_{21}$ | $Y_{22}$ | $Y_{23}$ | $Y_{24}$ | $Y_{25}$ |
| $Y_{31}$ | $Y_{32}$ | $Y_{33}$ | $Y_{34}$ | $Y_{35}$ |
| $Y_{41}$ | $Y_{42}$ | $Y_{43}$ | $Y_{44}$ | $Y_{45}$ |
| $Y_{51}$ | $Y_{52}$ | $Y_{53}$ | $Y_{54}$ | $Y_{55}$ |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| $\overline{Y}_{.1}$ | $\overline{Y}_{.2}$ | $\overline{Y}_{.3}$ | $\overline{Y}_{.4}$ | $\overline{Y}_{.5}$ |

$$Y_{ij} - \overline{Y}_{..} = (Y_{ij} - \overline{Y}_{.j}) + (\overline{Y}_{.j} - \overline{Y}_{..})$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij} - \overline{Y}_{.j})^2 + n\sum_{j=1}^{k}(\overline{Y}_{.j} - \overline{Y}_{..})^2$$

$$SS_{Total} = SS_{\text{Within Groups}} + SS_{\text{Between Groups}}$$

A sum of squares is the numerator of a variance. Dividing by the appropriate degrees of freedom, produces a variance estimate. In ANOVA, these variance estimates are called *mean squares*:

$$MS_{\text{Within Groups}} = \frac{SS_{\text{Within Groups}}}{N - k}$$

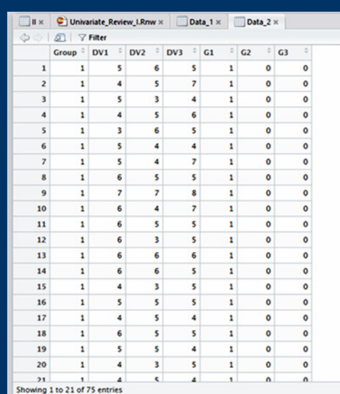
$$MS_{\text{Between Groups}} = \frac{SS_{\text{Between Groups}}}{k - 1}$$

$$F = \frac{MS_{\text{Between Groups}}}{MS_{\text{Within Groups}}}$$

If the null hypothesis is true, *F* has an expected value of approximately 1.00 (the numerator and denominator are estimates of the same variability).

If the null hypothesis is false, the numerator will be larger than the denominator because systematic, between-group differences contribute to the variance of the means, but not to the variance within groups (ideally). If *F* is "large enough," we will reject the null hypothesis as a reasonable description of the obtained variability.

In this example, there are 3 groups and 3 measures. The values on the measures can range from 1 to 10.

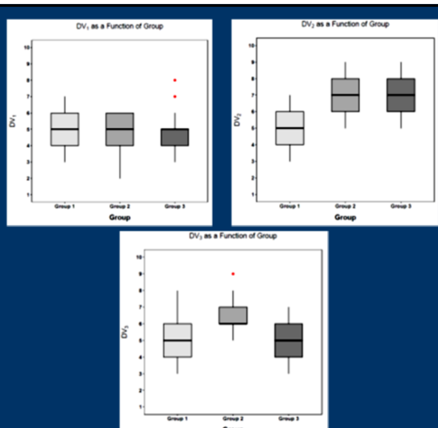


II × | Univariate_Review_I.Rnw × | Data_1 × | Data_2 ×

Filter

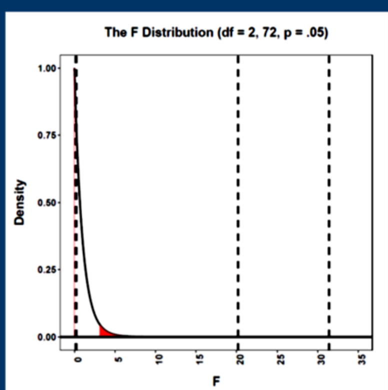
| | Group | DV1 | DV2 | DV3 | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 5 | 6 | 5 | 1 | 0 | 0 |
| 2 | 1 | 4 | 5 | 7 | 1 | 0 | 0 |
| 3 | 1 | 5 | 3 | 4 | 1 | 0 | 0 |
| 4 | 1 | 4 | 5 | 6 | 1 | 0 | 0 |
| 5 | 1 | 3 | 6 | 5 | 1 | 0 | 0 |
| 6 | 1 | 5 | 4 | 4 | 1 | 0 | 0 |
| 7 | 1 | 5 | 4 | 7 | 1 | 0 | 0 |
| 8 | 1 | 6 | 5 | 5 | 1 | 0 | 0 |
| 9 | 1 | 7 | 7 | 8 | 1 | 0 | 0 |
| 10 | 1 | 6 | 4 | 7 | 1 | 0 | 0 |
| 11 | 1 | 6 | 5 | 5 | 1 | 0 | 0 |
| 12 | 1 | 6 | 3 | 5 | 1 | 0 | 0 |
| 13 | 1 | 6 | 6 | 6 | 1 | 0 | 0 |
| 14 | 1 | 6 | 6 | 5 | 1 | 0 | 0 |
| 15 | 1 | 4 | 3 | 5 | 1 | 0 | 0 |
| 16 | 1 | 5 | 5 | 5 | 1 | 0 | 0 |
| 17 | 1 | 4 | 5 | 4 | 1 | 0 | 0 |
| 18 | 1 | 6 | 5 | 5 | 1 | 0 | 0 |
| 19 | 1 | 5 | 5 | 4 | 1 | 0 | 0 |
| 20 | 1 | 4 | 3 | 5 | 1 | 0 | 0 |
| 21 | 1 | 4 | 5 | 4 | 1 | 0 | 0 |

Showing 1 to 21 of 75 entries



A simple ANOVA can be obtained using the aov( ) function. Note that the group variable must be specified as a factor or it will be assumed to be a continuous predictor.

```
summary(aov(Data_2$DV1 ~ as.factor(Data_2$Group)))

##                         Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Data_2$Group)  2    0.4   0.213     0.2   0.82
## Residuals               72   78.2   1.087

summary(aov(Data_2$DV2 ~ as.factor(Data_2$Group)))

##                         Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(Data_2$Group)  2   72.1    36.1    31.4 1.5e-10
## Residuals               72   82.6     1.1

summary(aov(Data_2$DV3 ~ as.factor(Data_2$Group)))

##                         Df Sum Sq Mean Sq F value     Pr(>F)
## as.factor(Data_2$Group)  2   48.3    24.2    20.2 0.00000011
## Residuals               72   86.2     1.2
```




The sampling distribution for the ratio of two variance estimates is the F distribution.

ANOVA assumes homogeneity of within-group variances.

```
leveneTest(Data_2$DV1 ~ as.factor(Data_2$Group), data = Data_2)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2     0.1   0.91
##       72

leveneTest(Data_2$DV2 ~ as.factor(Data_2$Group), data = Data_2)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2       0      1
##       72

leveneTest(Data_2$DV3 ~ as.factor(Data_2$Group), data = Data_2)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2    0.14   0.87
##       72
```

Following a significant omnibus test, pairwise comparisons can be used to identify the nature of an effect.

```
pairwise.t.test(Data_2$DV2, as.factor(Data_2$Group), p.adj = "holm")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Data_2$DV2 and as.factor(Data_2$Group)
##
##   1     2
## 2 6e-09 -
## 3 6e-09 1
##
## P value adjustment method: holm
```

Multiple hypothesis tests increase the likelihood of falsely claiming a significant result (a Type I error). Correction procedures (of which there are a number) can control the error rate (e.g., FWER).

The F test, however, is a “blunt instrument.” It simply answers the question, “are the means different enough that their variability is larger than would be expected from the null hypothesis sampling distribution?”

More specific questions can be asked, using focused contrasts of groups (**L**) and variables (**M**, if there are repeated measures).

Maximum flexibility comes from using the L matrix approach to tailor the group comparisons.
To use this approach, we specify a "no intercept" model and then compare the resulting coefficients (which are group means).

A full set of dummy codes are needed for this analysis.

```
Data_2$G1 <- ifelse(Data_2$Group == 1, 1, 0)
Data_2$G2 <- ifelse(Data_2$Group == 2, 1, 0)
Data_2$G3 <- ifelse(Data_2$Group == 3, 1, 0)
```

```
DV1_LM <- lm(Data_2$DV1 ~ -1 + Data_2$G1 + Data_2$G2 + Data_2$G3]
DV2_LM <- lm(Data_2$DV2 ~ -1 + Data_2$G1 + Data_2$G2 + Data_2$G3]
DV3_LM <- lm(Data_2$DV3 ~ -1 + Data_2$G1 + Data_2$G2 + Data_2$G3]
```

Three regression coefficients are estimated in each model, but now there is no traditional intercept. Instead, each coefficient is a group mean and comparisons of the coefficients can be used to examine group differences.

The L matrix is applied to the coefficients and to the covariance matrix for the coefficients (to get the appropriate variance for the linear combination).

```
summary(DV2_LM)

##
## Call:
## lm(formula = Data_2$DV2 ~ -1 + Data_2$G1 + Data_2$G2 + Data_2$G3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.96  -0.92   0.04   1.04   2.12
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## Data_2$G1    4.880      0.214    22.8   <2e-16 ***
## Data_2$G2    6.960      0.214    32.5   <2e-16 ***
## Data_2$G3    6.960      0.214    32.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 72 degrees of freedom
## Multiple R-squared:  0.973,Adjusted R-squared:  0.972
## F-statistic:  877 on 3 and 72 DF,  p-value: <2e-16
```

The “estimates” are the group means.

Why are the standard errors precisely the same?

We can create any comparisons among the group estimates that we wish, including all of the pairwise comparisons. Each row is a different comparison.

```
L_Matrix <- matrix(c(1, -1, 0, 1, 0, -1, 0, 1, -1, 1, 1, -2, 2, -1,
    -1, 1, -2, 1), nrow = 6, ncol = 3, byrow = TRUE)
rownames(L_Matrix) <- c("G1 vs G2", "G1 vs G3", "G2 vs G3", "G1 and G2 vs G3",
    "G1 vs G2 and G3", "G2 vs G1 and G3")
L_Matrix

##                 [,1] [,2] [,3]
## G1 vs G2           1   -1    0
## G1 vs G3           1    0   -1
## G2 vs G3           0    1   -1
## G1 and G2 vs G3    1    1   -2
## G1 vs G2 and G3    2   -1   -1
## G2 vs G1 and G3    1   -2    1
```

The expected value of each comparison under the null hypothesis is 0.

```
glht_L_Matrix_2 <- glht(DV2_LM, linfct = L_Matrix, alternative = "two.sided",
    rhs = 0)
summary(glht_L_Matrix_2, adjusted("holm"))
```

```
  Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = Data_2$DV2 ~ -1 + Data_2$G1 + Data_2$G2 + Data_2$G3)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
G1 vs G2 == 0           -2.080      0.303   -6.87  9.7e-09
G1 vs G3 == 0           -2.080      0.303   -6.87  9.7e-09
G2 vs G3 == 0            0.000      0.303    0.00  1.00000
G1 and G2 vs G3 == 0    -2.080      0.525   -3.96  0.00051
G1 vs G2 and G3 == 0    -4.160      0.525   -7.93  1.2e-10
G2 vs G1 and G3 == 0    -2.080      0.525   -3.96  0.00051
(Adjusted p values reported -- holm method)
```

Unless comparisons are predicted by theory, we need to adjust for the number of comparisons being made.

```
V <- as.matrix(vcov(DV2_LM))
L1 <- as.matrix(L_Matrix[1, ])
V

##              Data_2$G1 Data_2$G2 Data_2$G3
## Data_2$G1     0.04587   0.00000   0.00000
## Data_2$G2     0.00000   0.04587   0.00000
## Data_2$G3     0.00000   0.00000   0.04587

L1

##       [,1]
## [1,]     1
## [2,]    -1
## [3,]     0

t(L1) %*% V %*% L1

##          [,1]
## [1,] 0.09173

sqrt(t(L1) %*% V %*% L1)

##         [,1]
## [1,] 0.3029
```

```
## Linear Hypotheses:
##                    Estimate Std. Error t value Pr(>|t|)
## G1 vs G2 == 0        -2.080      0.303   -6.87   <0.001 ***
```

Here is confirmation that the variance of the linear combination is obtained by applying the weights to the variance-covariance matrix of the variables being combined.

To this point, all of the examples have focused on a single measure at a time, effectively treating the M matrix as a scalar equal to 1.

Often we have more than one measure and have linear combinations of them that we wish to examine.

Next time . . .

Extending ANOVA to repeated measures and the use of the M matrix.