Inferential Statistics

semicolon

Learning Outcomes

- Participants should understand the theory and application of inferential statistics, Parameter estimation, Hypothesis testing, Analysis of Variance

Contingency Test

- If you have two categorical variables and you want to compare whether there is a relationship between two variables, you can use the contingency test
- The null hypothesis means that the two categorical variables have no relationship. The alternate hypothesis means that the two categorical variables have a relationship

Contingency Test

- To calculate the expected value, use

$$E_{ij} = \frac{R_i C_j}{N}$$

- where *R* is the row, *C* is the column, *N* is the total, *ith* is the row, and *jth* is the column.The formula for *X*2 statistics is
-

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Contingency Test

```
cht<-read.csv(file.choose(),header = T)
chisq.test(cht$color,cht$clarity)
Pearson's Chi-squared test

data:  cht$color and cht$clarity
X-squared = 2047.1, df = 42, p-value < 2.2e-16
```

Contingency Test

The chi-square test:

- *H*0: The two variables are independent.
- *Ha*: The two variables are not independent.

The p-value is 2.2e-16, so it is less than 0.05, which is the alpha value. Therefore, the null hypothesis is to be rejected. The two variables are not independent.

ANOVA

ANOVA is the process of testing the means of two or more groups. ANOVA also checks the impact of factors by comparing the means of different samples. In a t-test, you test the means of two samples; in a chi-square test, you test categorical attributes or variables; in ANOVA, you test means of two or more groups

Grand Mean

In ANOVA, you use two kinds of means, sample means and a grand mean. A grand mean is the mean of all of the samples' means.

Hypothesis

In ANOVA, a null hypothesis means that the sample means are equal or do not have significant differences. The alternate hypothesis is when the sample means are not equal

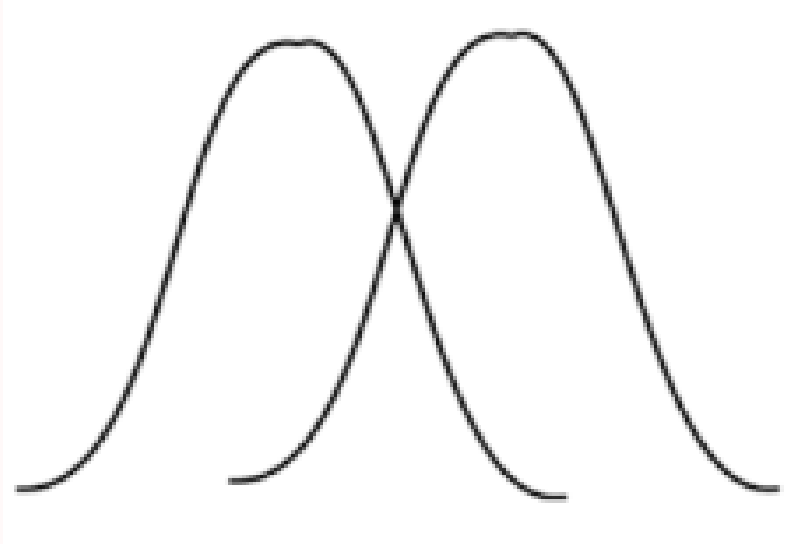
$H_0 : \mu_1 = \mu_2 = \ldots = \mu_L$ *Null hypothesis*

$H_a : \mu_1 \neq \mu_m$ *Alternate hypothesis*

Assumptions of Anova

- You assume that the variables are sampled, independent, and selected or sampled from a population that is normally distributed with unknown but equal variances
- **Between Group Variability**
- The distribution of two samples, when they overlap, their means are not significantly different. Hence, the difference between their individual mean and the grand mean is not significantly different. The group and level are different groups in the same independent variable

Between Group Variability



Means are not Significantly Different

Between Group Variability

- *Means are Significantly Different*

semicolon

Sum of Squares Between

To calculate the sum of the square of between the group variability, use

$$SS_{between} = n_1\left(\bar{x}_1 - \bar{x}_G\right)^2 + n_2\left(\bar{x}_2 - \bar{x}_G\right)^2 + n_3\left(\bar{x}_3 - \bar{x}_G\right)^2 + \ldots + n_k\left(\bar{x}_k - \bar{x}_k\right)^2$$

where
xG is the grand mean,
x*1… *xk is the mean of each sample, and
***n*1…*nk*…** are the sample sizes.
To calculate the sum of each squared deviation, or mean square, use

$$MS_{between} = \frac{n_1\left(\bar{x}_1 - \bar{x}_G\right)^2 + n_2\left(\bar{x}_2 - \bar{x}_G\right)^2 + n_3\left(\bar{x}_3 - \bar{x}_G\right)^2 + \ldots + n_k\left(\bar{x}_k - \bar{x}_k\right)^2}{k-1}$$

Sum of Squares Within

- You use the SS to divide by the degree of freedom, where the degree of freedom is the number of sample means(k) minus one

- **Within Group Variability**
- Within-group variation refers to the variations caused by differences within individual groups or levels. To calculate the sum of squares of within-group variation, use
- $$SS_{within} = \sum \left(x_{i1} - \bar{x}_1 \right)^2 + \sum \left(x_{i2} - \bar{x}_2 \right)^2 + \ldots + \sum \left(x_{ik} - \bar{x}_3 \right)^2 = \sum \left(x_{ij} - \bar{x}_j \right)^2$$

Sum of Square Cont'd

where
*xi*1 is the *ith* value of first sample,
*xi*2 is the *ith* value of second sample, and
xij is the *jth* value from the jth sample.
The degree of freedom is

$$df_{within} = (n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1) = n_1 + n_2 + n_3 + \ldots + n_k - k(1) = N - k$$

Mean Square (within)

- To get the mean square of the within-group variability, you divide between group variability sum of the squares with degree of freedom within:

$$MS_{within} = \sum \left(x_{ij} - \overline{x_j} \right)^2 / \left(N - k \right)$$

-

$$F - statistics = \frac{Between - group\ variability}{Within - group\ variablility} = \frac{MS_{between}}{MS_{within}}$$

If the f-critical value is smaller than the f-value, reject the null hypothesis. The f-critical value can be found using F-statistics and the degree of freedom on the f distribution

One-Way ANOVA

One-way ANOVA is used when you have only one independent variable.
In R, you can calculate the one-way ANOVA using

```
> a1<-aov(cht$price~cht$color)
> summary(a1)
              Df    Sum Sq   Mean Sq F value Pr(>F)
cht$color      6 2.685e+10 4.475e+09   290.2 <2e-16 ***
Residuals  53933 8.316e+11 1.542e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The p-value is less than 0.05, so we reject the null hypothesis and conclude that the mean price of diamond for the colors are not equal.

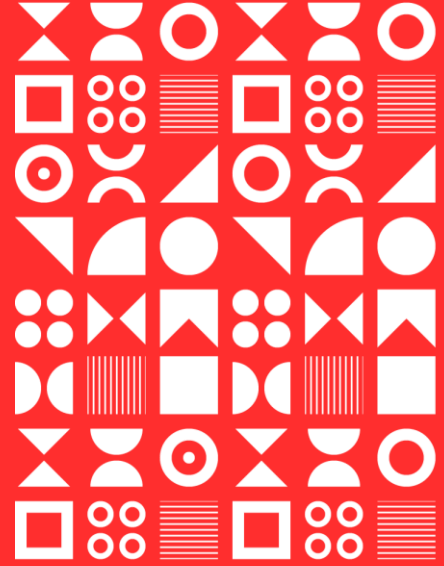
- Post Hoc Test
- We want to know which of the pairs of color are significantly different? Hence we can carry out post hoc test, there are numerous post hoc but in this lesson we will use turkeyhsd test.

```
> TukeyHSD(a1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = cht$price ~ cht$color)

$`cht$color`
             diff        lwr        upr     p adj
E-D   -93.20162 -276.13675   89.73351 0.7437450
F-D   554.93230  371.00057  738.86403 0.0000000
G-D   829.18158  651.26607 1007.09708 0.0000000
H-D  1316.71510 1127.17600 1506.25419 0.0000000
I-D  1921.92086 1710.95948 2132.88224 0.0000000
J-D  2153.86392 1894.02250 2413.70535 0.0000000
F-E   648.13392  481.61576  814.65208 0.0000000
G-E   922.38320  762.53532 1082.23107 0.0000000
H-E  1409.91672 1237.22484 1582.60860 0.0000000
I-E  2015.12248 1819.15789 2211.08707 0.0000000
J-E  2247.06554 1999.24508 2494.88601 0.0000000
G-F   274.24927  113.26182  435.23673 0.0000106
H-F   761.78280  588.03556  935.53004 0.0000000
I-F  1366.98856 1170.09331 1563.88381 0.0000000
J-F  1598.93162 1350.37458 1847.48867 0.0000000
H-G   487.53352  320.16821  654.89884 0.0000000
I-G  1092.73928  901.45210 1284.02646 0.0000000
J-G  1324.68235 1080.54376 1568.82093 0.0000000
I-H   605.20576  403.06243  807.34909 0.0000000
J-H   837.14882  584.41420 1089.88345 0.0000000
J-I   231.94307  -37.23053  501.11666 0.1449244
```

Reference

- All IT eBooks. “Learn R for Applied Statistics - PDF EBook Free Download.” *Allitebooks.In*, 13 Feb. 2019, www.allitebooks.in/learn-r-for-applied-statistics/.

Questions ??



semicolon