Analysis of Variance

Peter Ralph

6 October – Advanced Biological Statistics

Outline

Goal

To compare means of something between groups.

Goal

To compare means of something between groups.

Related topics:

- When can you do it, and how well? Power, false positive rate.
- How can experiments best do it? Experimental design.
- Methods: two-sample t-test, (one-way) ANOVA, permutation tests

Comparing means

Example:

How different are AirBnB prices between neighbourhoods?

```
airbnb <- read.csv("../Datasets/portland-airbnb-listings.csv", stringsAsFactors=TRUE)
airbnb$price <- as.numeric(gsub("$", "", airbnb$price, fixed=TRUE))
airbnb$neighbourhood[airbnb$neighbourhood == ""] <- NA
(neighbourhood_counts <- sort(table(airbnb$neighbourhood), decreasing=TRUE))
```

```
##
##            Richmond   Northwest District           Concordia            Downtown
##                 318                  238                 230                 221
##           Mt. Tabor            Irvington    Sellwood-Moreland          Montavilla
##                 145                  134                 133                 129
##               Cully        South Portland            Woodlawn           St. Johns
##                  87                   87                  85                  81
##   Creston-Kenilworth             Hillsdale   Old Town/Chinatown   Beaumont-Wilshire
##                  68                   59                  59                  58
##      University Park        Foster-Powell           Laurelhurst           Multnomah
##                  48                   47                  45                  45
##           Portsmouth              Alameda           Forest Park           Homestead
##                  37                   35                  34                  34
##            Hillside             Ashcreek      Pleasant Valley    Parkrose Heights
##                  25                   24                  21                  19
##       Lloyd District              Markham                Argay        Collins View
##                  11                   11                  10                  10
##        Hayden Island            Hollywood            Maplewood        Marshall Park
##                   7                    7                   7                   7
##        Woodland Park
##                   1                    0
```

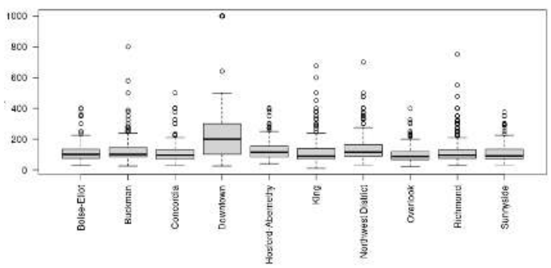
Let's take only the ten biggest neighbourhoods:

```
big_neighbourhoods <- names(neighbourhood_counts)[1:10]
sub_bnb <- subset(airbnb, !is.na(price) & neighbourhood %in% big_neighbourhoods)
sub_bnb <- droplevels(sub_bnb[, c("price", "neighbourhood", "host_id")])
nrow(sub_bnb)
```

```
## [1] 2023
```

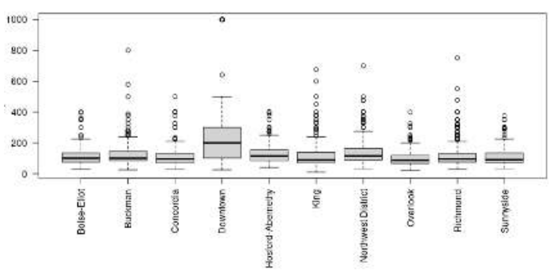

Look at the data:

```
par(mar=c(9, 3, 1, 1)+.1)
plot(price ~ neighbourhood, data=sub_bnb, fill=grey(0.8), las=2, xlab='')
```



Look at the data:

```
par(mar=c(9, 3, 1, 1)+.1)
plot(price ~ neighbourhood, data=sub_bnb, fill=grey(0.8), las=2, xlab='')
```



Preliminary conclusions? Formal questions?

ANOVA

The ANOVA model

The *price* P_{ij} of the jth room in neighbourhood i is

$$P_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where

- μ is the overall mean
- α_i is the mean deviation of neighborhood i from μ
- ϵ_{ij} is what's left over ("error", or "residual")

The ANOVA model

The *price* P_{ij} of the jth room in neighbourhood i is

$$P_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where

- μ is the overall mean
- α_i is the mean deviation of neighborhood i from μ
- ϵ_{ij} is what's left over ("error", or "residual")

In words,

$$(\text{price}) = (\text{group mean}) + (\text{residual})$$

ANOVA

- Stands for ANalysis Of VAriance
- Core statistical procedure in biology
- Developed by R.A. Fisher in the early 20th Century
- Core idea: ask how much variation exists within vs. among groups
- ANOVAs are linear models that have categorical predictor and continuous response variables
- The categorical predictors are often called factors, and can have two or more levels

Question 1: what are the means?

```r
summary(lm(formula = price ~ neighbourhood, data = sub_bnb))
```

```
##
## Call:
## lm(formula = price ~ neighbourhood, data = sub_bnb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -211.70  -48.12  -23.16   17.28  762.30
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     118.15625    8.13700  14.521   <2e-16 ***
## neighbourhoodBuckman             11.10976   10.88102   1.021   0.3074
## neighbourhoodConcordia           -5.53016   10.59577  -0.522   0.6018
## neighbourhoodDowntown           118.53940   10.83458  10.941   <2e-16 ***
## neighbourhoodHosford-Abernethy   14.56073   11.52553   1.263   0.2066
## neighbourhoodKing                 3.14322   11.08430   0.284   0.7768
## neighbourhoodNorthwest District  23.42358   10.52246   2.226   0.0261 *
## neighbourhoodOverlook           -13.53446   11.58099  -1.169   0.2427
## neighbourhoodRichmond            -0.03638    9.98145  -0.004   0.9971
## neighbourhoodSunnyside           -3.90324   11.40300  -0.342   0.7322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Question 2: is there group heterogeneity?

I.e.: do mean prices differ by neighborhood?

Question 2: is there group heterogeneity?

I.e.: do mean prices differ by neighborhood?

How would *you* do this?

Design a statistic that would be big if mean prices are different between neighborhoods, and will be small if all neighborhoods are the same.

Question 2, answered by ANOVA

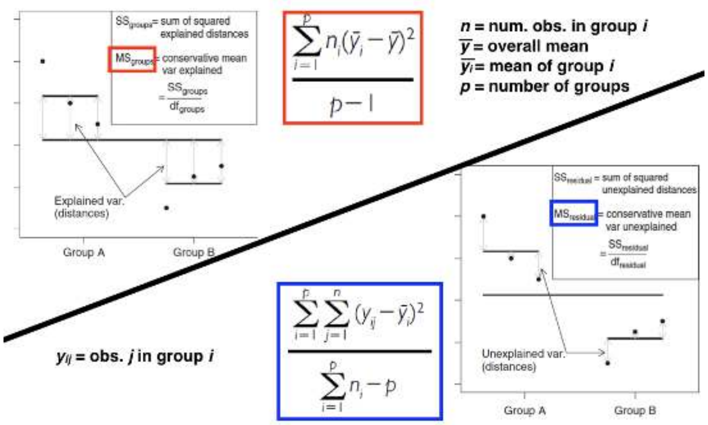
```
anova(lm(formula = price ~ neighbourhood, data = sub_bnb))
```

```
## Analysis of Variance Table
##
## Response: price
##                 Df   Sum Sq Mean Sq F value    Pr(>F)
## neighbourhood    9  2655967  295107  27.857 < 2.2e-16 ***
## Residuals     2013 21325161   10594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Table 8.2 ANOVA table for single factor linear model showing partitioning of variation

Source of	SS	df	MS	
Between groups	$\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2$	$p-1$	$\frac{\sum_{i=1}^{p} n_i(\bar{y}_i - \bar{y})^2}{p-1}$	**Var. explained by groupings**
Residual	$\sum_{i=1}^{p}\sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$	$\sum_{i=1}^{p} n_i - p$	$\frac{\sum_{i=1}^{p}\sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{p} n_i - p}$	**Var. unexplained by groupings**
Total	$\sum_{i=1}^{p}\sum_{j=1}^{n} (y_{ij} - \bar{y})^2$	$\sum_{i=1}^{p} n_i - 1$		

$$\text{F-ratio} = \frac{MS_{groups}}{MS_{residuals}}$$



One or more predictor variables

- One-way ANOVAs just have a single factor
- Multi-factor ANOVAs
  - Factorial - two or more factors and their interactions
  - Nested - the levels of one factor are contained within another level
  - The models can be quite complex
- ANOVAs use an F-statistic to test factors in a model
  - Ratio of two variances (numerator and denominator)
  - The numerator and denominator d.f. need to be included (e.g. $F_{1,34} = 29.43$)
- Determining the appropriate test ratios for complex ANOVAs takes some work

Assumptions

- Normally distributed groups
  - robust to non-normality if equal variances and sample sizes
- Equal variances across groups
  - okay if largest-to-smallest variance ratio < 3:1
  - problematic if there is a mean-variance relationship among groups
- Observations in a group are independent
  - randomly selected
  - don't confound group with another factor

// reveal.js plugins

/