Chapter 12

Analysis of Variance (ANOVA)

12.1 One-Way ANOVA

12.1.1 Comparing three or more populations

We have seen that the weight of a newborn baby is affected by whether or not the mother smokes cigarettes (see Example 7.9). Does the age of the mother also affect the birth weight of a baby? Figure 12.1 displays the distribution of birth weights of a random sample of boys born in Illinois in 2004 to mothers in the age ranges of 15-19, 20-24 and 25-29 years. [1] There appears to be a trend of weight increasing as age increases, but is this difference statistically significant?

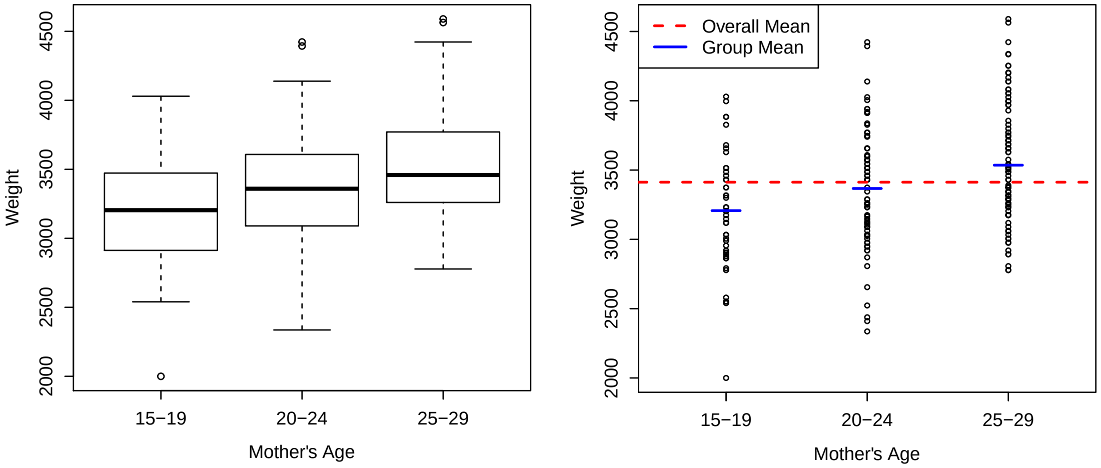

Figure 12.1: Distribution of the birth weights of boys born in Illinois in 2004. In the right figure, a horizontal line marks the overall mean, and a triangle indicates the age group means.

In Chapters 3, 7 and 8, we encountered methods for comparing two populations. In this section, we will compare the means of three or more populations. The classical approach is called the analysis of variance, or *ANOVA* for short. Contrary to the name given to this method, we will not be comparing the variances of our populations, but rather the variability in their means. ANOVA takes into account how the birth weights in each age group vary from the groups mean, and how the weights vary between the age groups, using the information in Table 12.1.

[1] The births are also restricted to single births only and gestation lengths of at least 37 weeks

Age	15-19	20-24	25-29
mean	3207.205	3367.000	3535.363
sd	422.003	410.682	413.589
n	44	90	107

Table 12.1: Summary statistics for birth weights.

12.1.2 The ANOVA F Test

The ANOVA F test is for comparing means of populations.

Assume we have independent random samples drawn from G groups (populations). In the birth weight example the $G = 3$ populations are baby boys born to mothers in each of the three age groups. In a medical experiment investigating weight loss under different treatments, the populations might represent obese individuals on the Atkins diet, a vegan diet, a high carbohydrate diet, and a control group (G=4).

Let μ_g the true mean in group g, $g = 1, 2, \ldots, G$. The hypothesis of interest is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_G$$

versus

$$H_A : \mu_i \neq \mu_j \text{ for some } i \neq j.$$

Suppose there are n_g observations in the sample from the g^{th} group and $n = n_1 + n_2 + \cdots + n_G$. Let Y_{gk} denote the response of the k^{th} observation in the g^{th} group.

Group	Observations	Group mean
1	Y_{11} Y_{12} $\cdots$ Y_{1n_1}	$\bar{Y}_{1.}$
2	Y_{21} Y_{22} $\cdots$ Y_{2n_2}	$\bar{Y}_{2.}$
$\vdots$	$\vdots$	$\vdots$
G	Y_{G1} Y_{G2} $\cdots$ Y_{Gn_G}	$\bar{Y}_{G.}$

Table 12.2: Observations drawn from the G populations.

We can describe the model using the *cell means model*,

$$Y_{gk} = \mu_g + \epsilon_{gk},$$

with all Y's independent, $\mathrm{E}[Y_{gk}] = \mu_g$ and $\mathrm{Var}[Y_{gk}] = \sigma^2$, $k = 1, 2, \ldots, n_g$, $g = 1, 2, \ldots, G$. The ϵ_{gk} represents random error; it follows that $\mathrm{E}[\epsilon_{gk}] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2$.

Let $\bar{Y}g.$ denote the mean for the sample from the g^{th} group,

$$\bar{Y}_{g.} = \frac{1}{n_g}\sum_{k=1}^{n_g} Y_{gk},$$

and $\bar{Y}_{..}$ denote the overall sample mean (often called the grand mean),

$$\bar{Y}_{..} = \frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n_g} Y_{gk} = \frac{1}{n}\sum_{g=1}^{G} n_g \bar{Y}_{g.}.$$

The idea is to compare the variability between each group to the variability within each group. Thus, for variability between the groups, we look at $(\bar{Y}_{g.} - \bar{Y}_{..})$, the amount that the group means differ from the overall mean. For within group variability we look at $(Y_{gk} - \bar{Y}_{g.})$, that is, the amount that the n_g individuals in the g^{th} group differ from that group's mean. Thus, if the means μ_g, $g = 1, 2, \ldots, G$, are all the same, then the variability between the groups and the variability within the groups should be roughly the same.

We now define the *treatment sum of squares*,

$$SSTR = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g.} - \bar{Y}_{..})^2 = \sum_{g=1}^{G} n_g(\bar{Y}_{g.} - \bar{Y}_{..})^2,$$

the *error sum of squares*,

$$SSE = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g.})^2,$$

and the *total sum of squares*,

$$SST = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{..})^2,$$

the last sum representing the variability of each observation from the overall mean. We can partition this overall variability (SST) into the between group variability plus the within group variability.

Theorem 12.1. $SST = SSTR + SSE$

Proof.

$$\begin{aligned}
SST &= \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{..})^2 \\
&= \sum_{g=1}^{G}\sum_{k=1}^{n_g}\left[(\bar{Y}_{g.} - \bar{Y}_{..}) + (Y_{gk} - \bar{Y}_{g.})\right]^2 \\
&= \sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g.} - \bar{Y}_{..})^2 + 2\sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g.} - \bar{Y}_{..})(Y_{gk} - \bar{Y}_{g.}) + \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g.})^2 \\
&= \sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g.} - \bar{Y}_{..})^2 + 0 + \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g.})^2
\end{aligned}$$

The proof that

$$\sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g.} - \bar{Y}_{..})(Y_{gk} - \bar{Y}_{g.}) = 0$$

is left as an exercise. □

Remark We have seen the idea of partitioning the variability of a quantity before in the context of least-squares regression, Section 9.3.2. ||

The sample variance of the observations in group g is

$$S_g^2 = \frac{1}{n_g - 1}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g.})^2,$$

so the error sum of squares can be expressed as

$$SSE = \sum_{g=1}^{G} (n_g - 1) S_g^2.$$

We can pool these estimates of the sample variances across all groups

$$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \cdots (n_G-1)S_G^2}{(n_1-1)+(n_2-1)+\cdots+(n_G-1)} = \frac{\sum_{g=1}^{G}(n_g-1)S_g^2}{n-G} = \frac{SSE}{n-G} \tag{12.1}$$

to obtain an unbiased estimate of the common variance σ^2.

Theorem 12.2. *Let Y_{gk}, $k = 1, 2, \ldots, n_g$, $g = 1, 2, \ldots, G$, denote independent random variables with* $\mathrm{E}[Y_{gk}] = \mu_g$, $\mathrm{Var}[Y_{gk}] = \sigma^2$. *Let* $\mu = \sum_{g=1}^{G} n_g \mu_g$.

Then

$$\mathrm{E}\Big[\frac{SSTR}{G-1}\Big] = \sigma^2 + \frac{1}{G-1}\sum_{g=1}^{G} n_g(\mu_g - \mu)^2 \tag{12.2}$$

$$\mathrm{E}\Big[\frac{SSE}{n-G}\Big] = \sigma^2 \tag{12.3}$$

Proof. We prove Equation 12.2; the proof of Equation 12.3 follows similar reasoning.

First, it is easy to check that $\mathrm{E}[\bar{Y}_{..}] = \mu$ and $\mathrm{Var}[\bar{Y}_{..}] = \sigma^2/n$. Thus, by Theorem A.5, $\mathrm{E}[(\bar{Y}_{..} - \mu)^2] = \mathrm{Var}[\bar{Y}_{..}] = \sigma^2/n$.

In addition, $\mathrm{Var}[(\bar{Y}_{g.} - \mu)] = \mathrm{E}[(\bar{Y}_{g.} - \mu)^2] - (\mathrm{E}[(\bar{Y}_{g.} - \mu)])^2$, by Proposition A.2. Thus, since $\mathrm{Var}[\bar{Y}_{g.} - \mu)] = \mathrm{Var}[\bar{Y}_{g.}] = \sigma^2/n_g$, we have $\sigma^2/n_g = \mathrm{E}[\bar{Y}_{g.} - \mu)^2] - (\mathrm{E}[(\bar{Y}_{g.} - \mu)])^2$, or rewriting,

$$\mathrm{E}[(\bar{Y}_{g.} - \mu)^2] = \sigma^2/n_g + (\mu_g - \mu)^2.$$

Thus,

$$\begin{aligned}
\mathrm{E}[SSTR/(G-1)] &= \frac{1}{G-1}\mathrm{E}\Big[\sum_{g=1}^{G} n_g(\bar{Y}_{g.} - \bar{Y}_{..})^2\Big] \\
&= \frac{1}{G-1}\mathrm{E}\Big[\sum_{g=1}^{G} n_g\big[(\bar{Y}_{g.} - \mu) - (\bar{Y}_{..} - \mu)\big]^2\Big] \\
&= \frac{1}{G-1}\mathrm{E}\Big[\sum_{g=1}^{G} n_g\big[(\bar{Y}_{..} - \mu)^2 - 2(\bar{Y}_{g.} - \mu)(\bar{Y}_{..} - \mu) + (\bar{Y}_{..} - \mu)^2\big]\Big] \\
&= \frac{1}{G-1}\mathrm{E}\Big[\sum_{g=1}^{G} n_g(\bar{Y}_{g.} - \mu)^2 - 2(\bar{Y}_{..} - \mu)\sum_{g=1}^{G} n_g(\bar{Y}_{g.} - \mu) + \sum_{g=1}^{G} n_g(\bar{Y}_{..} - \mu)^2\Big] \\
&= \frac{1}{G-1}\mathrm{E}\Big[\sum_{g=1}^{G} n_g(\bar{Y}_{g.} - \mu)^2 - 2(\bar{Y}_{..} - \mu)n(\bar{Y}_{..} - \mu) + n(\bar{Y}_{..} - \mu)^2\Big] \\
&= \frac{1}{G-1}\mathrm{E}\Big[\sum_{g=1}^{G} n_g(\bar{Y}_{g.} - \mu)^2\Big] - n\frac{1}{G-1}\mathrm{E}[(\bar{Y}_{..} - \mu)^2] \\
&= \frac{1}{G-1}\sum_{g=1}^{G} n_g\big(\sigma^2/n_g + (\mu_g - \mu)^2\big) - \frac{1}{G-1}n\sigma^2/n
\end{aligned}$$

$$= \sigma^2 + \frac{1}{G-1}\sum_{g=1}^{G} n_g(\mu_g - \mu)^2$$

□

Thus, if the population means are all the same, $\mu_1 = \mu_2 = \cdots = \mu_G = \mu$, then the sum in Equation 12.2 is zero so on average, the ratio $\dfrac{SSTR/(G-1)}{SSE/(n-G)}$ should be equal to 1. Otherwise, if $(\mu_g - \mu)^2 > 0$ for at least one g, then on average the ratio is greater than 1.

If we also assume that the populations are normally distributed, we can say more.

Theorem 12.3. *Let Y_{gk}, $k = 1, 2, \ldots, n_g$, $g = 1, 2, \cdots, G$, denote independent random variables, $Y_{gk} \sim N(\mu_g, \sigma^2)$.*

Then

1. *SSE and $SSTR$ are independent.*
2. *SSE/σ^2 has a chi-square distribution with $n - G$ degrees of freedom.*
3. *If $\mu_1 = \mu_2 = \cdots = \mu_G$, then $SSTR/\sigma^2$ has a chi-square distribution with $G - 1$ degrees of freedom.*

Proof. The proofs of parts (1) and (2) are left as exercises, and the proof of part (3) is omitted. □

Definition 12.1 The treatment sum of squares divided by its degrees of freedom is called the *mean square for treatments*,

$$MSTR = \frac{SSTR}{G-1}$$

The error sum of squares divided by its degrees of freedom is called the *mean squared error* (or *mean squared residual*),

$$MSE = \frac{SSE}{n-G}. \tag{12.4}$$

||

Theorem 12.4. *Let Y_{gk}, $k = 1, 2, \ldots, n_g$, $g = 1, 2, \cdots, G$, denote independent random variables, $Y_{gk} \sim N(\mu_g, \sigma^2)$, If $\mu_1 = \mu_2 = \cdots = \mu_G$, then*

$$F = \frac{MSTR}{MSE}$$

has an F distribution with $G - 1$ and $n - G$ degrees of freedom.

Proof.

$$F = \frac{MSTR}{MSE} = \frac{SSTR/(G-1)}{SSE/(n-G)} = \frac{\dfrac{(SSTR/\sigma^2)}{G-1}}{\dfrac{(SSE/\sigma^2)}{n-G}}. \tag{12.5}$$

By Theorem 12.3, $SSTR$ and SSE are independent, and $SSTR/\sigma^2$ and SSE/σ^2 have chi-square distributions with degrees of freedom $G - 1$ and $n - G$, respectively. Thus, the result follows by Definition B.9. □

Example 12.1 For the Illinois boys, the overall mean birth weight is $\bar{Y}_{..} = 3412.564$ grams. The treatment sum of squares is the sum (over observations) of the squared difference between the group mean for the observation and the overall mean (refer to Table 12.1). This simplifies to a sum (over groups) of the group size times the squared difference difference between group mean and overall mean:

$$\begin{aligned} SSTR &= \sum_{g=1}^{3}\sum_{k=1}^{n_g}(\bar{Y}_{g.} - 3412.564)^2 \\ &= \sum_{k=1}^{44}(3207.205 - 3412.564)^2 + \sum_{k=1}^{90}(3367.000 - 3412.564)^2 + \sum_{k=1}^{107}(3535.336 - 3412.6)^2 \\ &= 44(3207.205 - 3412.564)^2 + 90(3367.000 - 3412.564)^2 + 107(3535.336 - 3412.6)^2 \\ &= 3655236. \end{aligned}$$

The degrees of freedom for $SSTR$ is $G - 1 = 3 - 1 = 2$. Thus, the mean square for treatment is $3655236/2 = 1827628$.

The error sum of squares is the sum of squared difference between each observation and its group mean:

$$\begin{aligned} SSE &= \sum_{g=1}^{3}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}g.)^2 \\ &= \sum_{k=1}^{44}(Y_{k1} - 3207.205)^2 + \sum_{k=1}^{90}(Y_{k2} - 3367.00)^2 + \sum_{k=1}^{107}(Y_{k3} - 3535.336)^2 \\ &= 40800419. \end{aligned}$$

The degrees of freedom is $n - G = 241 - 3 = 239$, and the mean square for error is $40800419/238 = 171430.3$.

Thus, the F statistic is $1827628/171430.3 = 10.661$ and is compared to an F distribution with 2 and 238 degrees of freedom. The corresponding P-value is approximately 0.00004 so we conclude that the true means between the age groups are indeed different.

The calculations can be summarized in an ANOVA table.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mother's Age	2	3655256	1827628	10.661	0.00004
Residuals	238	40800419	171430		
Total	240	44455675			

Table 12.3: ANOVA Table

R Note:

The `lm` command that we used for linear regression in Chapter 9 used in conjunction with the `anova` command can be used to perform the ANOVA F test; or we can combine `summary` and `aov`.

The data for this example are in the file `ILBoys`.

```
> anova(lm(Weight ~ MothersAge, data = ILBoys))
             Df   Sum Sq  Mean Sq F value    Pr(>F)
MothersAge    2  3655256  1827628  10.661 3.679e-05 ***
Residuals   238 40800419   171430
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(lm(Weight ~ MothersAge, data = ILBoys))$F[1]    #Extract F statistic
[1] 10.66105


> summary(aov(Weight ~ MothersAge, data = ILBoys)) #same
...
```

□

Example 12.2 Measurements of the resistivity (ohms-cm) of silicon wafers were made at the National Institute of Standards and Technology (NIST) with 5 instruments on each of 5 days. Each of the 25 observations is the average of 6 measurements. (`http://www.itl.nist.gov/div898/strd/anova/SiRstv_info.html`).

If we let μ_i, $i = 1, 2, \ldots, 5$ denote the mean of the resistivity measurements for each of the instruments, then we wish to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_5 \text{ versus } H_A : \mu_i = \mu_j \text{ for some } i \neq j.$$

The overall mean resistance is 196.1892 ohm-cm with standard deviation 0.1056 ohm-cm.

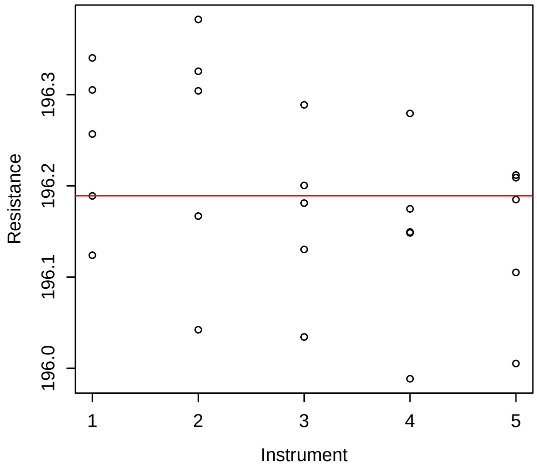

Figure 12.2: Distribution of the resistivity measurements across the five instruments. The horizontal line marks the overall mean.

The treatment sum of squares is $SSTR = 0.051146$ on 4 degrees of freedom, so the mean square for treatments is $MSTR = 0.051146/4 = 0.012787$.

The residual sum of squares is $SSE = 0.216637$ on $25 - 5 = 20$ degrees of freedom, so the mean square for error is $MSE = 0.216637/20 = 0.010832$.

Thus, the F statistic is $0.012787/0.010832 = 1.1805$ and we compare this to an F distribution with 4 and 20 degrees of freedom. The resulting P-value is 0.35 so we have no reason to reject the null hypothesis. We conclude that the mean resistance levels are the same across instruments. □

12.1.2.1 Assumptions

The assumptions needed to use the F tests parallel those described in Section 9.4 for linear regression, except for the assumption of linearity.

In other words, we assume:

- which group an observation falls into is fixed, not random,
- the residuals are all independent,
- the residuals have constant variance,
- the residuals are normally distributed.

Of those assumptions, the independence assumption is critical, and the others usually less important. The reasoning is similar to that in Section 9.4.3, but two assumptions merit extra comments here, the assumptions of constant variance and of normal distributions.

We motivate both comments by considering the case $G = 2$. Using ANOVA with only two groups is equivalent to doing a two-sample pooled-variance two-sided t-test. The t-test is preferred; it is easier to understand, allows for one-sided tests, and does not require pooling the variances.

The big problem with non-normality in t-tests is the effect of skewness on one-sided tests. But ANOVA tests are inherently two-sided (we are testing for *any* differences between means, not differences in one direction) so non-normal distributions generally have little effect as long as the sample sizes are reasonably large.

With ANOVA we are forced to assume that variances are equal. If the sample sizes n_g are roughly equal, then unequal variances don't hurt much, but if the population variances differ, then the actual sampling distribution could be very different from an F distribution. In particular, if there is a small sample from a population with large variance, then the F statistic can explode. To see this, consider an extreme case, where one sample is size 1, so the observation from that sample has no effect on SSE or MSE. If the variance for that population is a billion times the other population variances, then the $SSTR$ will tend to be much larger than we'd expect based on the MSE. With less extreme situations the effects will be more subtle, but the actual Type I error rate could be substantially different than the nominal value.

12.1.3 A Permutation Test Approach

The ANOVA test (Theorem 12.4) requires that the samples be drawn from normal populations. In addition, we must assume that the variances in the populations are the same. However, we can use the ideas encountered in Chapter 3 to form a permutation distribution of an appropriate test statistic. We randomly assign the values of the numeric variable to the k groups and compute the corresponding F statistic. We then note how extreme the observed F statistic is in the permutation distribution of the F statistic.

R **Note:**

```
observed <- anova(lm(Weight ~ MothersAge))$F[1]
n <- length(ILBoys$Weight)
B <- 10^5 - 1
results <- numeric(B)
for (i in 1:B)
{
  index <- sample(n)
  Weight.perm <- Weight[index]
  results[i] <- anova(lm(Weight ~ MothersAge))$F[1]
}

(sum(results> observed) + 1) / (B + 1)   # P value
```

On a slower computer, you may want to change B to a smaller number, say $B = 10^4 - 1$.

12.2 Exercises

1. In Theorem 12.1, prove $\sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g.} - \bar{Y}_{..})(Y_{gk} - \bar{Y}_{g.}) = 0$.

2. Prove parts (1) and (2) to Theorem 12.3.

3. In the early 1900's, Latter (1902) investigated the behavior of female cuckoos, that lay their eggs on the ground and then move them to the nests of other birds. In particular, Latter gathered data on the lengths of the cuckoo eggs found in these foster-nests. Data based on this work is used in (Tippett (1952)) and is located in the file `cuckoos`. The data contains the lengths, in millimeters, of the lengths of cuckoo eggs and the species of the nests where the eggs were placed.

   (a) Create side-by-side boxplots to compare the distribution of lengths across the different foster nests.

   (b) Conduct an ANOVA test to see if the mean lengths of the cuckoo eggs are the same across the different foster nests.

4. The file `wafers.csv` contains the data for Example 12.2.

   (a) Use `R` to replicate the results in this example.

   The `Instrument` values are numeric so in the `lm` command, you will need to use `as.factor(Instrument)` so that `R` treats this variable as a factor variable.

   (b) Conduct a permutation test to determine whether or not mean resistance measurements are the same across instruments.

5. Recall the Flight Delays Case Study 1.1 and the distribution of flight delay times for United Airlines across days (see Figure 2.6). Use a permutation test to determine whether or not the mean delay times across days of the week are the same.

6. Starcraft is a popular strategy video game with a science fiction military theme. Players choose to be one of three races—the Terrans, Zergs or Protoss'— and compete for dominance in a distance part of the Milky Way galaxy. The file `Starcraft` contains information on a sample of the top Korean players from the database `http://www.teamliquid.net/tlpd/players/` (Evans (2008)). In addition to the player's chosen race, the file contains his age as well as the number of games won (out of his most recent 40 games).

   (a) Use a permutation test to determine whether or not the mean age of the players is the same across races.

   (b) Use a permutation test to determine whether or not the mean number of wins is the same across races.