13 Chi-square- and F- tests

13.1 Review of χ^2-distribution (P.389-390)

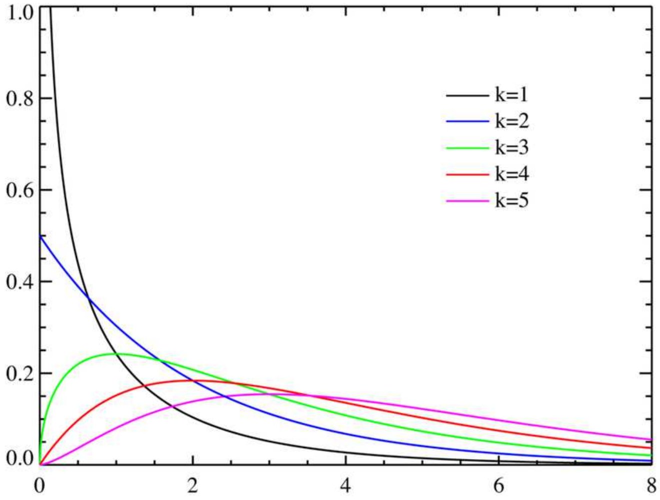
Definitions: If X is a standard normal random variable, the distribution of $U = X^2$ is called the *chi-square* distribution with degree of freedom 1.

If $U_1, U_2, ..., U_n$ are independent chi-square random variables with degree 1, the distribution of $V = U_1 + U_2 + ... + U_n$ is called the chi-square distribution with degree of freedom (df) n, and is denoted by $V \sim \chi^2_n$.

Density function of χ^2_n is

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x \geq 0.$$

with mean n and variance $2n$.



The density function of χ^2 with different degrees of freedom.

For df ≥ 3, the density funtions are *right-skewed*.

13.2 Review of F-distribution (P.391-394)

Definitions: Let U and V be independent chi-squared variables with m and n degrees of freedom respectively. The distribution of

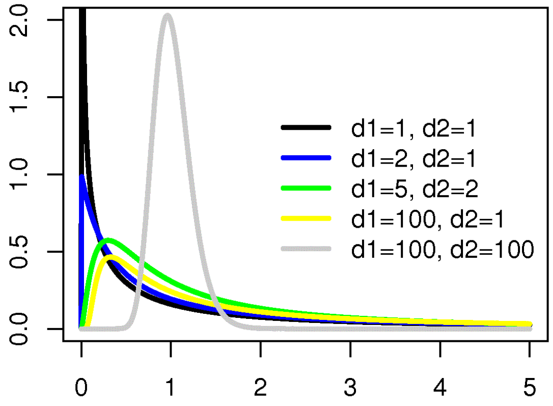
$$W = \frac{U/m}{V/n}$$

is called the *F-distribution* with m and n degrees of freedom and is denoted by $W \sim F_{m,n}$.

The density function of F-distribution is

$$f(x) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\,\Gamma(n/2)}(\frac{m}{n})^{m/2} x^{m/2-1}(1+\frac{m}{n}x)^{-(m+n)/2}, \quad x \geq 0.$$

with mean $\dfrac{n}{n-2}$ and variance $\dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, $n \geq 5$.



The density function of F with different degrees of freedom.

The R codes to calculate the values related to χ^2_n, i.e. $\Pr(\chi^2_n \leq q) = p$ and $F_{m,n}$, i.e. $\Pr(F_{m,n} \leq q) = p$ are

```
dchisq(x,n),  pchisq(q,n),  qchisq(p,n),  rchisq(v,n);
df(x,m,n),    pf(q,m,n),    qf(p,m,n),    rf(v,m,n).
```

where `d` gives density function $f(x)$, `p` gives lower area, `q` gives quantile and `r` gives a vector of values stored in `v` that follows the distribution.

Note that

$$F_{m,n} =_d \frac{1}{F_{n,m}}.$$

The sign '$=_d$' means 'equal in distribution'. Hence we have that, for any $x > 0$,

$$\Pr(F_{m,n} \geq x) = \Pr\left(\frac{U/m}{V/n} \geq x\right) = \Pr\left(\frac{V/n}{U/m} \leq \frac{1}{x}\right) = 1 - \Pr(F_{n,m} \geq 1/x).$$

That means in taking *reciprocal* of a ratio, one should *swap the dfs* and *change the tailed area* (from lower to upper or upper to lower).

13.3 Distributions derived from normal

Let X_i be iid random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$. The sample mean and sample variance are

$$\bar{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i \text{ and } S_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(X_i - \bar{X})^2.$$

Then the distributions derived from normal include:

1. $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n_x}} \sim \mathcal{N}(0, 1)$, a standard normal random variable.
2. $\dfrac{\bar{X} - \mu}{S_x/\sqrt{n_x}} \sim t_{n_x - 1}$, a Student t variable with $n - 1$ df.
3. $(n_x - 1)S_x^2/\sigma^2 = \sum_{i=1}^{n_x}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n_x - 1}$, a chi-squared variable with $n_x - 1$ df.

Note: $\sum_{i=1}^{n_x}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_{n_x}$ but $\sum_{i=1}^{n_x}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n_x - 1}$ loses 1 df because μ is estimated by $\bar{X}$.

Let $Y_i, 1 \le i \le n_y$, be iid normal random variables and $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$ independent of $X_i, 1 \le i \le n_x$. Writing

$$\bar{Y} = \frac{1}{n_y}\sum_{i=1}^{n_y} Y_i \text{ and } S_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(Y_i - \bar{Y})^2,$$

$$F = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} = \frac{(n_x - 1)S_x^2/[\sigma_x^2(n_x - 1)]}{(n_y - 1)S_y^2/[\sigma_y^2(n_y - 1)]} = \frac{\chi^2_{n_x - 1}/(n_x - 1)}{\chi^2_{n_y - 1}/(n_y - 1)} \sim F_{n_x - 1, n_y - 1}.$$

Note: in comparing two *scales* (σ^2; $H_0 : \sigma^2 = \sigma_0^2$ or $H_0 : \sigma_x^2 = \sigma_y^2$), one takes a *ratio* of sample variances, i.e. s^2/σ_0^2 or s_x^2/s_y^2, while in comparing two *locations* (μ), one takes a *difference* of sample means, i.e. $\bar{x} - \mu_0$ or $\bar{x} - \bar{y}$.

13.4 One sample Chi-square test for variance (P.410-418)

Example: (Variance of diameter) A company claims that the variance of the diameter of its machined engine parts is less than 0.0002 (measured in inches). A random sample of 10 parts gave a sample variance of 0.00018. Can we accept the company's claim?

Solution: Let X be the measured diameter. If the measured diameters are normally distributed, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$, then the question becomes to test:

$$H_0: \ \sigma^2 = 0.0002 \quad \text{vs} \quad H_1: \sigma^2 < 0.0002.$$

Suppose $X_1, X_2, ..., X_n$ is the sample and the mean μ and variance σ^2 are both unknown. The five steps test on the variance is

1. **Hypothesis:** $H_0: \sigma^2 = \sigma_0^2$ vs $H_1: \sigma^2 > \sigma_0^2,\ \sigma^2 < \sigma_0^2,\ \sigma^2 \neq \sigma_0^2$
2. **Test statistic:** $\chi_0^2 = (n-1)s^2/\sigma_0^2.$
3. **Assumption:** $X_i \sim \mathcal{N}(\mu, \sigma_0^2)$. Then $\chi_0^2 \sim \chi_{n-1}^2$ under H_0 and should be close to $n-1$.
4. ***P*-value:** $\Pr(\chi_{n-1}^2 \geq \chi_0^2)$ for $H_1: \sigma^2 > \sigma_0^2$,
   $\Pr(\chi_{n-1}^2 \leq \chi_0^2)$ for $H_1: \sigma^2 < \sigma_0^2$,
   $2\min\left\{\Pr(\chi_{n-1}^2 \geq \chi_0^2),\ \Pr(\chi_{n-1}^2 \leq \chi_0^2)\right\}$ for $H_1: \sigma^2 \neq \sigma_0^2$
5. **Decision:** reject H_0 if p-value $< \alpha$.

Note: $\chi_0^2 \gg n-1$ and $\chi_0^2 \ll n-1$ argue against H_0 in favor of $H_1: \sigma^2 > \sigma_0^2$ and $H_1: \sigma^2 < \sigma_0^2$ respectively.

13.5 Rejection region and confidence interval

Since the χ^2 distribution in the upper and lower sides are

$$\chi^2_{n-1,1-\alpha} \gg n-1 \quad \text{and} \quad \chi^2_{n-1,\alpha} \ll n-1 \;\; \text{respectively,}$$

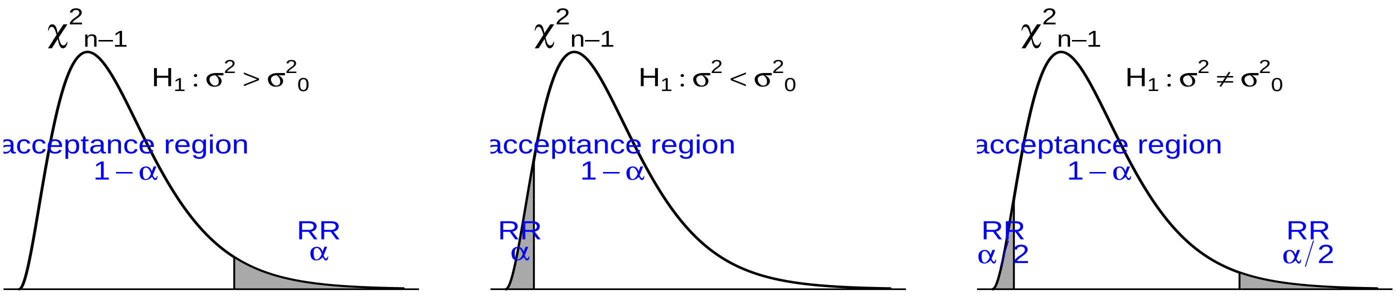
the ratios in the upper and lower sides are

$$\frac{\chi^2_{n-1,1-\alpha}}{n-1} \gg 1 \quad \text{and} \quad \frac{\chi^2_{n-1,\alpha}}{n-1} \ll 1 \;\; \text{respectively.}$$

Hence given a significance level α, the *rejection region* for the sample variance s^2 in which we should reject H_0 and accepted H_1 is

$$\boxed{\begin{array}{lll} s^2 \;\geq\; \sigma_0^2 \dfrac{\chi^2_{n-1,1-\alpha}}{n-1} & & \text{for } H_1 : \sigma^2 > \sigma_0^2; \\ s^2 \;\leq\; \sigma_0^2 \dfrac{\chi^2_{n-1,\alpha}}{n-1} & & \text{for } H_1 : \sigma^2 < \sigma_0^2; \\ s^2 \;\leq\; \sigma_0^2 \dfrac{\chi^2_{n-1,\alpha/2}}{n-1} \quad \text{or} \quad s^2 \geq \sigma_0^2 \dfrac{\chi^2_{n-1,1-\alpha/2}}{n-1} & & \text{for } H_1 : \sigma^2 \neq \sigma_0^2, \end{array}}$$

where $\Pr(\chi^2_{n-1} \leq \chi^2_\alpha) = \alpha$.



Test on χ_0^2	$\chi^2_{n-1,1-\alpha}$	$\chi^2_{n-1,\alpha}$	$\chi^2_{n-1,\alpha/2}$	$\chi^2_{n-1,1-\alpha/2}$
Test on s^2	$\sigma_0^2\frac{\chi^2_{n-1,1-\alpha}}{n-1}$	$\sigma_0^2\frac{\chi^2_{n-1,\alpha}}{n-1}$	$\sigma_0^2\frac{\chi^2_{n-1,\alpha/2}}{n-1}$	$\sigma_0^2\frac{\chi^2_{n-1,1-\alpha/2}}{n-1}$

Moreover the confidence intervals for testing σ_0^2 are

$$\left(s^2\frac{n-1}{\chi^2_{n-1,1-\alpha}},\ \infty\right) \qquad \text{for } H_1:\sigma^2>\sigma_0^2;$$

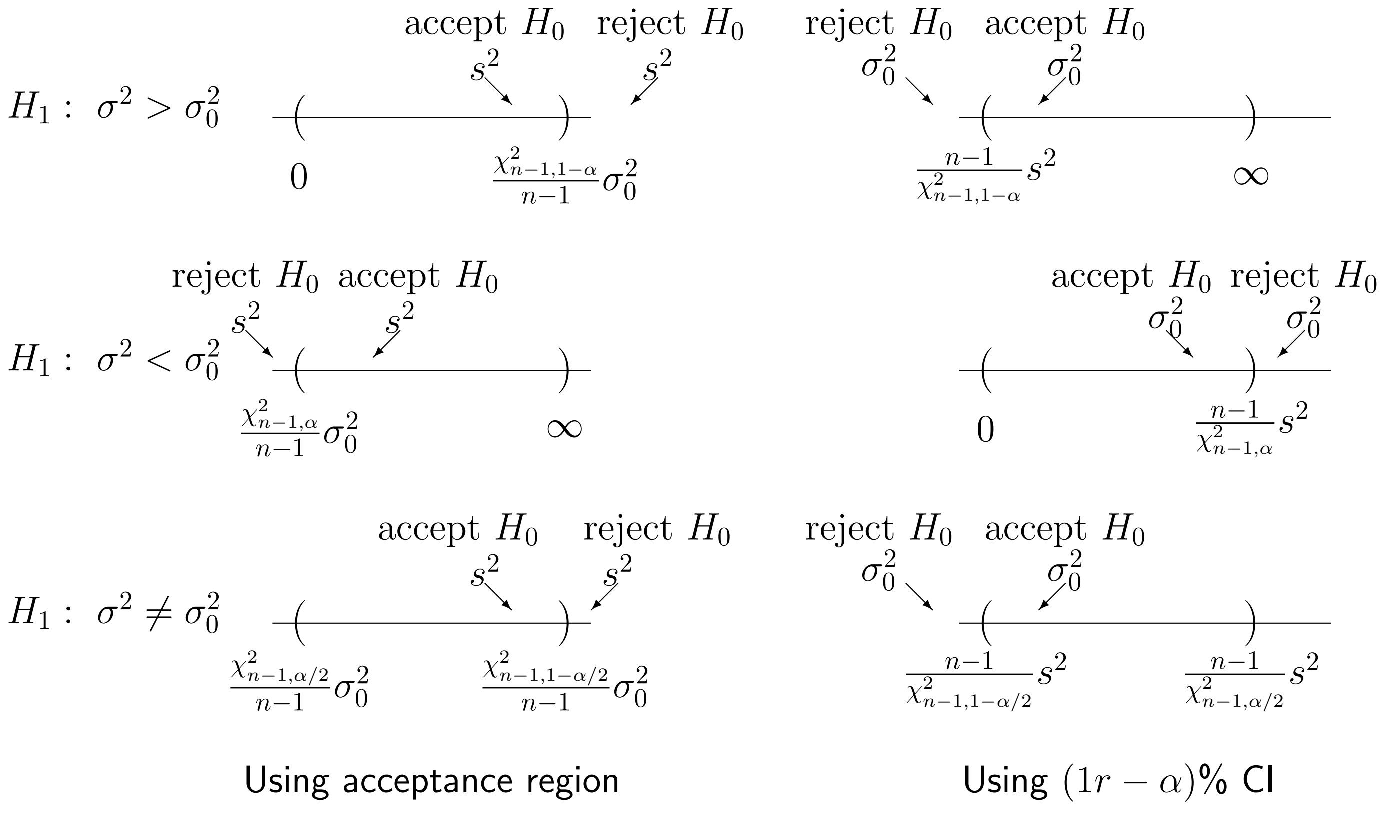
$$\left(0,\ s^2\frac{n-1}{\chi^2_{n-1,\alpha}}\right) \qquad \text{for } H_1:\sigma^2<\sigma_0^2;$$

$$\left(s^2\frac{n-1}{\chi^2_{n-1,1-\alpha/2}},\ s^2\frac{n-1}{\chi^2_{n-1,\alpha/2}}\right) \qquad \text{for } H_1:\sigma^2\neq\sigma_0^2.$$

Note that the reciprocal of ratio $\dfrac{n-1}{\chi^2_{n-1,1-\alpha}}$ is used.

If σ_0^2 lies in the CI, we accept H_0. Otherwise, we reject H_0.

In summary,



$H_1:\ \sigma^2>\sigma_0^2$

accept H_0 reject H_0

s^2 s^2

0 $\quad \frac{\chi^2_{n-1,1-\alpha}}{n-1}\sigma_0^2$

reject H_0 accept H_0

σ_0^2 σ_0^2

$\frac{n-1}{\chi^2_{n-1,1-\alpha}}s^2$ $\quad \infty$

$H_1:\ \sigma^2<\sigma_0^2$

reject H_0 accept H_0

s^2 s^2

$\frac{\chi^2_{n-1,\alpha}}{n-1}\sigma_0^2$ $\quad \infty$

accept H_0 reject H_0

σ_0^2 σ_0^2

0 $\quad \frac{n-1}{\chi^2_{n-1,\alpha}}s^2$

$H_1:\ \sigma^2\neq\sigma_0^2$

accept H_0 reject H_0

s^2 s^2

$\frac{\chi^2_{n-1,\alpha/2}}{n-1}\sigma_0^2$ $\quad \frac{\chi^2_{n-1,1-\alpha/2}}{n-1}\sigma_0^2$

reject H_0 accept H_0

σ_0^2 σ_0^2

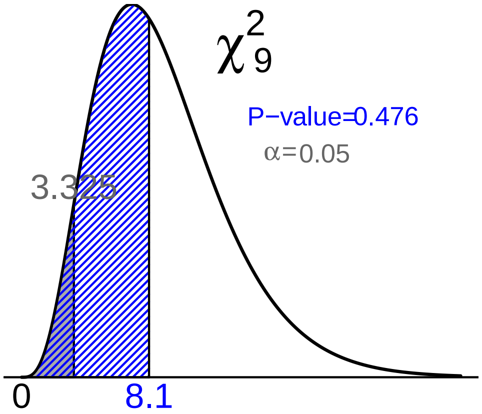
$\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}s^2$ $\quad \frac{n-1}{\chi^2_{n-1,\alpha/2}}s^2$

Using acceptance region

Using $(1r-\alpha)$% CI

Example: (Variance of diameter) The Chi-square test for the variance of the measurements is

1. **Hypothesis:** $H_0: \ \sigma^2 = 0.0002 \quad \text{vs} \quad H_1 : \sigma^2 < 0.0002.$
2. **Test statistic:** $\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9\times.00018}{.0002} = 8.1$
3. **Assumption:** $X_i \sim \mathcal{N}(\mu, 0.0002)$ under H_0. Then $\chi_0^2 \sim \chi_{n-1}^2$.
4. ***P*-value:** p-value $= \Pr(\chi_9^2 \leq 8.1) = 0.4759.$
5. **Decision:** Since the p-value > 0.05, the data is consistent with H_0 that the variance is 0.0002.



The median and mean are 8.3428 and 9 respectively.

In R,

```
> n=10
> sigma20=0.0002
> s2=0.00018
> chi20=(n-1)*s2/sigma20
> chi20
[1] 8.1
> p.value=pchisq(chi20,n-1)
> p.value
```

```
[1] 0.4758991
> alp=0.05
> infty=9999999 #represent infinity
> CI.upper=c((n-1)*s2/qchisq(1-alp,n-1),infty) #H1: sigma^2>sigma0^2
> CI.upper
[1] 9.575047e-05 9.999999e+06
> CI.lower=c(0,(n-1)*s2/qchisq(alp,n-1)) #H1: sig^2<sig0^2, this example
> CI.lower
[1] 0.000000  4.872015e-04
> CI.2sided=c((n-1)*s2/qchisq(1-alp/2,n-1),(n-1)*s2/qchisq(alp/2,n-1))
> CI.2sided
[1] 8.516111e-05 5.999135e-04
```

Since the lower 95% CI (0, 0.000487) includes the hypothesized variance of 0.0002, the data are consistent with $H_0 : \sigma^2 = 0.0002$.

13.6 Two samples F-test on variances (P.471-476)

Suppose that the samples $X_1, X_2, ..., X_{n_1}$ and $Y_1, Y_2, ..., Y_{n_2}$ are from two distinct normal populations with unknown variances σ_x^2 and σ_y^2 respectively. We want to test the hypothesis:

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad \text{vs} \quad H_1 : \ \sigma_x^2 > \sigma_y^2, \quad \text{or} \quad \sigma_x^2 \neq \sigma_y^2.$$

The sample variances S_x^2 and S_y^2 are estimates of σ_x^2 and σ_y^2 respectively. If H_0 holds, then S_x^2/S_y^2 should be close to 1. Then the variable

$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{n_x-1,n_y-1}$$

follows a F- distribution with mean $(n_y-1)/(n_y-3) > 1$ and degrees of freedom $n_x - 1$ and $n_y - 1$.

The five steps for the F-test are

1. **Hypotheses:** $H_0 : \ \sigma_x^2 = \sigma_y^2$ vs $H_1 : \ \sigma_x^2 > \sigma_y^2$ or $\sigma_x^2 \neq \sigma_y^2$
2. **Test statistic:** $f_0 = s_x^2/s_y^2 \quad \text{if} \quad s_x^2 > s_y^2.$
3. **Assumption:** $X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$ and X_i & Y_i are independent. Then $f_0 \sim F_{n_x-1,n_y-1}$ under H_0 and should be close to 1.
4. ***P*-value:** $\Pr(F_{n_x-1,n_y-1} \geq f_0)$ for $H_1 : \sigma_x^2 > \sigma_y^2$

   $2\Pr(F_{n_x-1,n_y-1} \geq f_0)$ for $H_1 : \sigma_x^2 \neq \sigma_y^2$
5. **Decision:** Reject H_0 if p-value $< \alpha$.

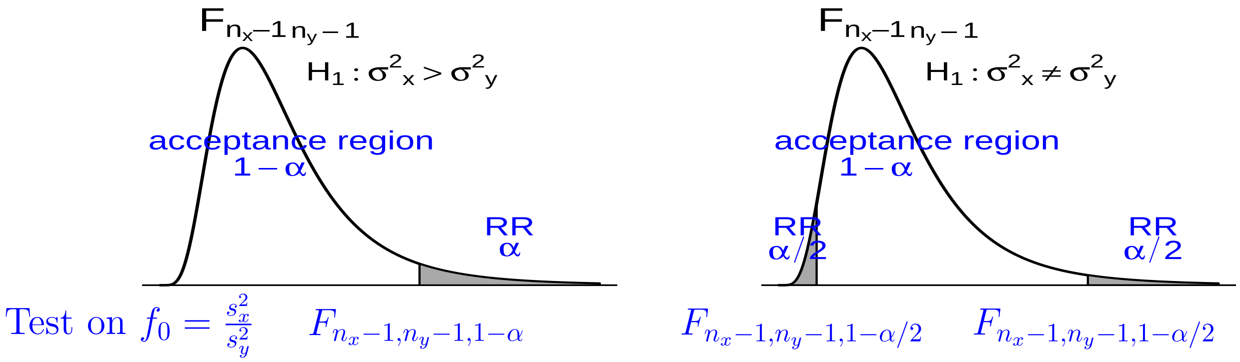
Note: $f_0 \gg 1$ or $f_0 \ll 1$ argue against H_0 in favor of $H_1 : \sigma_x^2 > \sigma_y^2$ and $H_1 : \sigma_x^2 < \sigma_y^2$ respectively.

13.7 Rejection region and confidence interval (P.483-484, 486-487)

Given significant level α, the rejection regions for the ratio of sample variances s_x^2/s_y^2 in which we should reject H_0 and accepted H_1 are

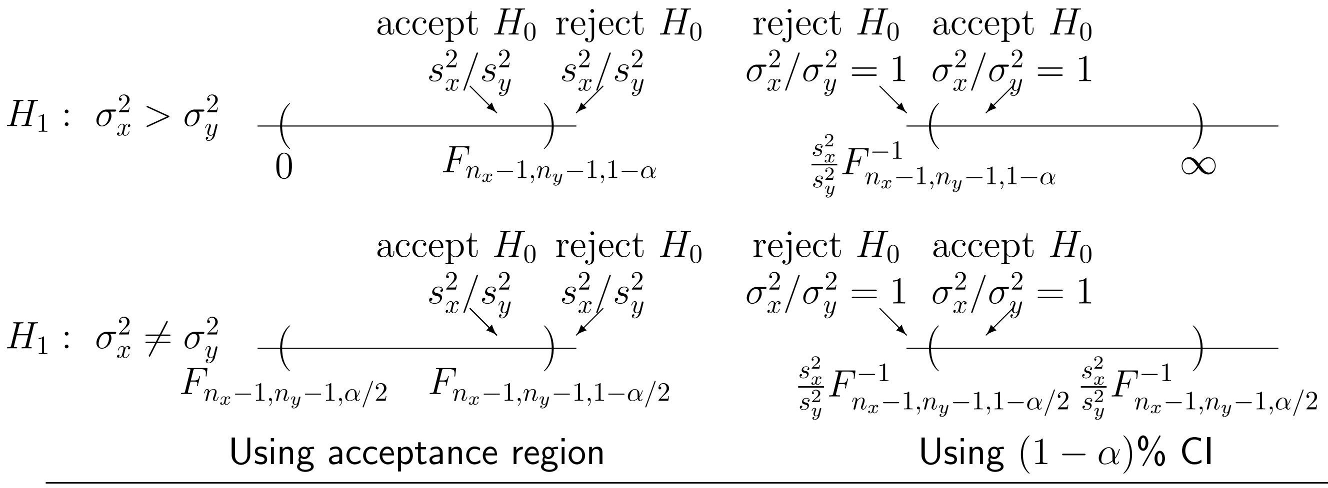
$$s_x^2/s_y^2 \geq F_{n_x-1,n_y-1,1-\alpha} \qquad \text{for } H_1\colon \sigma_x^2 > \sigma_y^2;$$
$$s_x^2/s_y^2 \leq F_{n_x-1,n_y-1,\alpha/2} \text{ or } s_x^2/s_y^2 \geq F_{n_x-1,n_y-1,1-\alpha/2} \quad \text{for } H_1\colon \sigma_x^2 \neq \sigma_y^2,$$

where $\Pr(F_{n_x-1,n_y-1} \leq F_\alpha) = \alpha$.



The $(1-\alpha)\%$ CIs for the ratio of true variances σ_x^2/σ_y^2 are

$$\left((s_x^2/s_y^2)\, F^{-1}_{n_x-1,n_y-1,1-\alpha},\ \infty\right) \qquad \text{for } H_1 : \sigma_x^2 > \sigma_y^2;$$
$$\left((s_x^2/s_y^2)\, F^{-1}_{n_x-1,n_y-1,1-\alpha/2},\ (s_x^2/s_y^2)\, F^{-1}_{n_x-1,n_y-1,\alpha/2}\right) \text{ for } H_1 : \sigma_x^2 \neq \sigma_y^2,$$



Example: (Height comparison) A fourth grade class has 10 girls and 13 boys. The children's heights are recorded on their 10th birthday as follows:

```
Boys:  135.3,  137.0, 136.0, 139.7, 136.5, 137.2, 138.8,
       139.6,  140.0, 137.7, 135.5, 134.9, 139.5
Girls: 140.3,  139.8, 138.6, 137.1, 140.0, 136.2,
       138.7,  138.5, 134.9, 141.0
```

Do the data provide sufficient evidence to indicate a difference in the variability of heights for boys and girls?

Solution: We have $n_x = 13$ and $n_y = 10$. Assume that the boys' height $X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and the girls' height $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$. The F test is

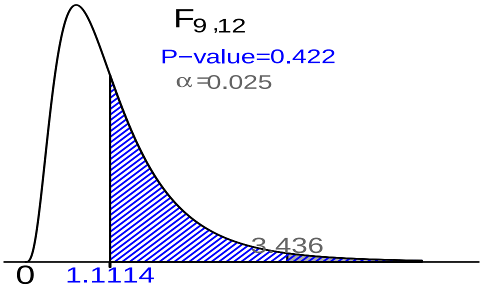
1. **Hypotheses:** $H_0: \ \sigma_x^2 = \sigma_y^2 \quad \text{vs } H_1: \ \sigma_x^2 \neq \sigma_y^2$
2. **Test statistic:** $f_0 = s_y^2/s_x^2 = \dfrac{3.7432}{3.3681} = 1.1114.$
3. **Assumption:** $X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$ and X_i & Y_i are independent. Then $f_0 \sim F_{n_1-1,n_2-1}$ under H_0.
4. ***P*-value:**

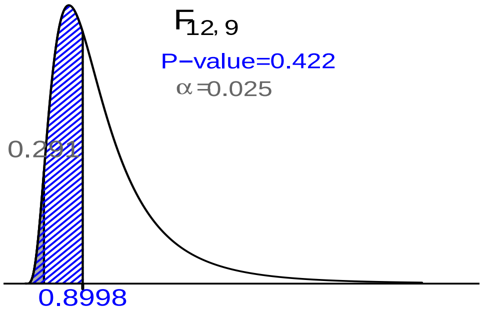
$$\begin{aligned}
& 2\Pr(F_{9,12} \geq 1.1114) \text{ or } 2\Pr(F_{12,9} \leq 1/1.1114) \\
= \ & 2\Pr(F_{9,12} \geq 1.1114) \\
= \ & 2(0.4224) = 0.8448 \quad \text{(from R)} \\
> \ & 2(0.1) = 0.2 \quad (\text{from F table, } F_{9,12,0.9} = 2.21)
\end{aligned}$$

5. **Decision:** The data is consistent with H_0. There is no evidence to indicate a difference in the variability of heights for boys and girls.

In R,

```
> x=c(135.3,137.0,136.0,139.7,136.5,137.2,138.8,139.6,140.0,137.7,135.5,
      134.9,139.5)
> y=c(140.3,139.8,138.6,137.1,140.0,136.2,138.7,138.5,134.9,141.0)
> var.test(y,x)

        F test to compare two variances

data:  y and x
F = 1.1114, num df = 9, denom df = 12, p-value = 0.8448
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.323467 4.299073
sample estimates:
ratio of variances
          1.111383

> n1=length(y)  #checking only
> n2=length(x)
> c(var(y),var(x),n1,n2)
[1]  3.743222  3.368077 10.000000 13.000000
> f0=var(y)/var(x)
> f01=1/f0
> p.value=2*(1-pf(f0,n1-1,n2-1)) #or p.val=2*pf(f0,n1-1,n2-1,lower.tail=F)
> p.value1=2*pf(1/f0,n2-1,n1-1)
> c(f0,f01,p.value,p.value1) #same p-value
[1] 1.1113826 0.8997801 0.8447837 0.8447837
> alp=0.05
> infty=9999999
> CI.upper=c(var(y)/(var(x)*qf(1-alp,n1-1,n2-1)),infty) #H1: sigy^2>sigx^2
> CI.upper
[1] 3.974368e-01 9.999999e+06
> CI.2sided=c(var(y)/(var(x)*qf(1-alp/2,n1-1,n2-1)),
             var(y)/(var(x)*qf(alp/2,n1-1,n2-1)))
> CI.2sided
[1] 0.323467 4.299073
```





Median of F(9,12) = 0.98 and median of F(12,9) = 1.0194.

Notes: These tests depend very strongly on assumptions of normality–unlike tests for mean.

14 Experimental design (P.619-621)

In a two-sample test, how should we design the experiments to obtain a more effective test?

14.1 Factors affecting the power

Consider the *large sample* z-test at a significant level α for the hypotheses:

$$H_0: \ \mu = \mu_0 \quad \text{vs} \quad H_1: \ \mu \neq \mu_0,$$

we accept H_0 if μ_0 lies inside the $100(1-\alpha)\%$ confidence interval (CI)

$$[\bar{X} - z_{\alpha/2}S/\sqrt{n}, \ \bar{X} + z_{\alpha/2}S/\sqrt{n}].$$

The power of the test for μ_1 in H_1 is

$$\text{Power}(\mu_1) = \Pr(\text{Reject } H_0 | \ \mu = \mu_1) = \Pr(\mu_0 \text{ lies outside the CI}).$$

Hence the higher the power of the z test, the shorter should be the CI. This happens when

1. the data have less variation, measured by S and
2. the sample size n increases.

If it is not possible to increase the sample size (like experiments very costly in money or time), the accuracy of a test may still be improved by *scientific design* (data collection) that reduces the variation of data.

In a two-sample test, two questions raise in the data collection:

1. For the totally n observations, what should the allocation to each sample be? If $n = 10$, should we select $n_x = n_y = 5$ or $n_x = 4$ and $n_y = 6$?
2. For a two-sample test, should we select paired samples or independent samples?

14.2 Allocation to increase accuracy

Example: Let $X_1, X_2, ..., X_{n_x}$ be a random sample from a population X with mean μ_x and variance σ_x^2, and $Y_1, Y_2, ..., Y_{n_y}$ be a random sample from another population Y with mean μ_y and variance σ_y^2.

If $n_x + n_y = n$, then

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y},$$

which is minimized at

$$\boxed{n_x = \frac{\sigma_x}{\sigma_x + \sigma_y} n, \quad n_y = \frac{\sigma_y}{\sigma_x + \sigma_y} n.}$$

In particular, if $\sigma_x^2 = \sigma_y^2$, then $n_x = n_y = n/2$.

Proof. Recall that if X and Y are independent, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

Hence,

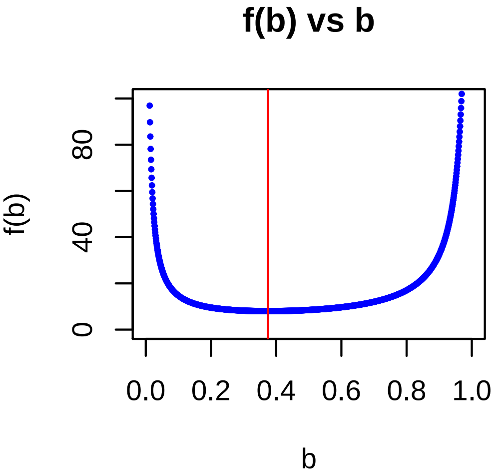
$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}.$$

Let b be the proportion of the n observations assigned to the sample from population X; that is, $n_x = nb$ and $n_y = (1-b)n$. Then,

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{bn} + \frac{\sigma_y^2}{(1-b)n}.$$

To find the proportion b that minimizes this variance, we set the first derivative, with respect to b, equal to zero. This process yields

$$\frac{\sigma_x^2}{n}(-1)\frac{1}{b^2}+\frac{\sigma_y^2}{n}(-1)\frac{1}{(1-b)^2}(-1)=0$$
$$\Rightarrow\ \frac{\sigma_y^2}{(1-b)^2}=\frac{\sigma_x^2}{b^2}\Rightarrow\frac{\sigma_y}{(1-b)}=\frac{\sigma_x}{b}$$
$$\Rightarrow\ b\sigma_y=(1-b)\sigma_x\Rightarrow b(\sigma_x+\sigma_y)=\sigma_x$$
$$\Rightarrow\ b=\frac{\sigma_x}{\sigma_x+\sigma_y}$$



since $\dfrac{dx^m}{dx}=mx^{m-1}$ and $\dfrac{d(1-b)^{-1}}{db}=\dfrac{d(1-b)^{-1}}{d(1-b)}\times\dfrac{d(1-b)}{db}$. Hence we obtain

$$b=\frac{\sigma_x}{\sigma_x+\sigma_y}\quad\text{and}\quad 1-b=\frac{\sigma_y}{\sigma_x+\sigma_y}.$$

Thus, $Var(\bar{X}-\bar{Y})$ is minimized when

$$n_x=\frac{\sigma_x}{\sigma_x+\sigma_y}n\quad\text{and}\quad n_y=\frac{\sigma_y}{\sigma_x+\sigma_y}n$$

such that more information (higher sample size) is needed when the population has higher variability (higher σ).

In particular, if $\sigma_x^2=\sigma_y^2$, then $n/2$ observations should be taken from each population.

Example: Suppose that you wish to compare the means for two populations with variances $\sigma_x^2 = 9$ and $\sigma_y^2 = 25$ respectively.

Find the smallest sample size and the corresponding sample allocation that will yield a 95% confidence interval for $\mu_x - \mu_y$ that is *two units or less* in length.

Solution: We obtain that $n_x = \frac{3}{3+5}n = \frac{3n}{8}$ and $n_y = \frac{5}{3+5}n = \frac{5n}{8}$, where n is the combined sample size. Since the true σ_x^2 and σ_y^2 are given, the 95% confidence interval for $\mu_x - \mu_y$ is given by

$$\left((\bar{X} - \bar{Y}) - z_{0.025}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}},\ (\bar{X} - \bar{Y}) + z_{0.025}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right)$$

where $z_{0.025} = 1.96$ and it is denoted by: $(\underline{\quad}, \underline{\quad})$.

We choose the sample size n such that the width of the CI is 2, i.e.

$$2 \times 1.96\sqrt{\frac{9}{\frac{3n}{8}} + \frac{25}{\frac{5n}{8}}} = 2 \times 1.96\sqrt{\frac{24}{n} + \frac{40}{n}} = 2$$

$$\Rightarrow\ 1.96\sqrt{\frac{64}{n}} = 1 \quad \Rightarrow\ \sqrt{\frac{n}{64}} = 1.96$$

$$\Rightarrow\ n = 1.96^2 \times 64 = 245.9 \approx 246$$

This implies that

$$\begin{aligned} n_x &= \frac{3}{8}(246) = 92.25 \approx 93, \\ n_y &= \frac{5}{8}(246) = 153.73 \approx 154. \end{aligned}$$

Note that the sample sizes are rounded *up* in order to ensure that *CI is not more than 2 units in length.*

14.3 Paired sample or independent sample? (P.649-652)

Suppose we want to compare two competing medications (A and B). An experiment can be designed in different ways for comparison.

For the *completely randomized* design, we recruit n individuals and randomly split them into two groups who are given medication A and B respectively. This leads to *independent* samples.

Alternatively, for the *match paired (block)* design, we recruit *pairs* of individuals of similar weight and age and same gender. One individual from each pair is *randomly selected* to receive medication A whereas the other receives medication B. This leads to *paired* samples.

Which experiment is better for the purpose of comparison?
Match paired sample gives more precise estimate because paired samples decrease the variability of data:

$$Var_m(\bar{D}) = Var(\bar{X} - \bar{Y}) = \frac{1}{n}(\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y) < \frac{1}{n}(\sigma_x^2 + \sigma_y^2) = Var_i(\bar{X} - \bar{Y}),$$

when the correlation coefficient is positive ($\rho > 0$). Hence if it is reasonable to believe that the pairs (X_j, Y_j) will tend to increase or decrease together in values ($\rho > 0$), use match paired sample.

For the model assumptions,

two independent samples test requires *normality* and *equality of variance* assumptions for *both populations* when the sample sizes are small

whereas

matched paired test requires only *normality for the differences* and hence is *less restrictive in distribution assumptions.* Moreover the matched pairs reduce the inherent variability in the paired difference.

However, it is *more restrictive on recruiting matched-pair subjects* than to recruit subjects for two independent samples.

15 One way ANOVA test I

What if we want to compare 3 or more means?

15.1 Introduction (P.430-455, 595-599)

The test is performed on the collection of samples taken independently from each of the g populations. The data are

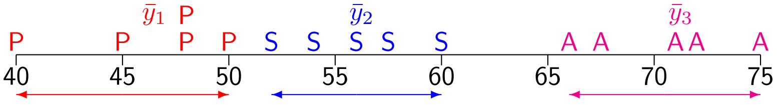
Treatment	1	2	$\cdots$	g	Grand total
	Y_{11} (Y_{11}^2) Y_{12} (Y_{12}^2) $\vdots$ $\vdots$ Y_{1n_1} $(Y_{1n_1}^2)$	Y_{21} (Y_{21}^2) Y_{22} (Y_{22}^2) $\vdots$ $\vdots$ Y_{2n_2} $(Y_{2n_2}^2)$	$\cdots$ $\cdots$ $\vdots$ $\cdots$	Y_{g1} (Y_{g1}^2) Y_{g2} (Y_{g2}^2) $\vdots$ $\vdots$ Y_{gn_g} $(Y_{gn_g}^2)$	
Column total Sample size	$Y_{1\cdot}$ $(\sum_j Y_{1j}^2)$ n_1	$Y_{2\cdot}$ $(\sum_j Y_{2j}^2)$ n_2	$\cdots$	$Y_{g\cdot}$ $(\sum_j Y_{gj}^2)$ n_3	$Y_{\cdot\cdot}$ $(\sum_{ij} Y_{ij}^2)$ N
Sample mean Sample var.	$\overline{Y}_1$ S_1^2	$\overline{Y}_2$ S_2^2	$\cdots$ $\cdots$	$\overline{Y}_g$ S_g^2	$\overline{Y}$ S^2

To compare the $g > 2$ population means, we test the hypotheses:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_g \quad \text{vs}$$
$$H_1: \text{Not all means equal.}$$

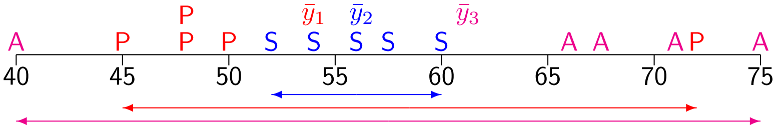
Example 1: *Same total sum of squares* SST_0 from the same set of y_{ij} to split into $g = 3$ groups:

Treatment Means Are Significantly Different



Same SST_0 (same data)
Larger SST (clear gp or more diff. $\bar{y}_i$)
Smaller SSR (less spread or smaller s_i^2)
Larger F (sign. diff.)

Treatment Means Are Not Significantly Different



Same SST_0 (same data)
Smaller SST (merge gp or less diff. $\bar{y}_i$)
larger SSR (wide spread or larger s_i^2)
Smaller F (no diff.)

In the one-way *Analysis of Variance* (ANOVA) test, we compare two sources of variance to make decision on whether the population means are equal. The following example illustrates the idea.

Example 2: Same *sum of squares for treatments* SST: same group means

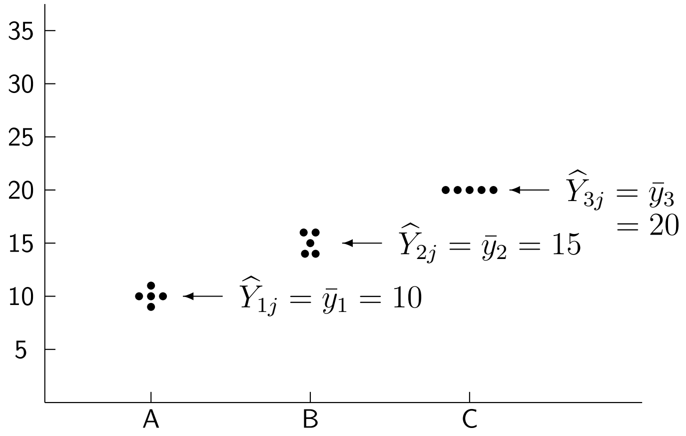
Small MSR: variation within gp
Large $F = \frac{MST}{MSR}$ & reject H_0

Groups		
10	15	20
10	16	20
11	14	20
10	16	20
9	14	20
$\bar{y}_1 = 10$	$\bar{y}_2 = 15$	$\bar{y}_3 = 20$

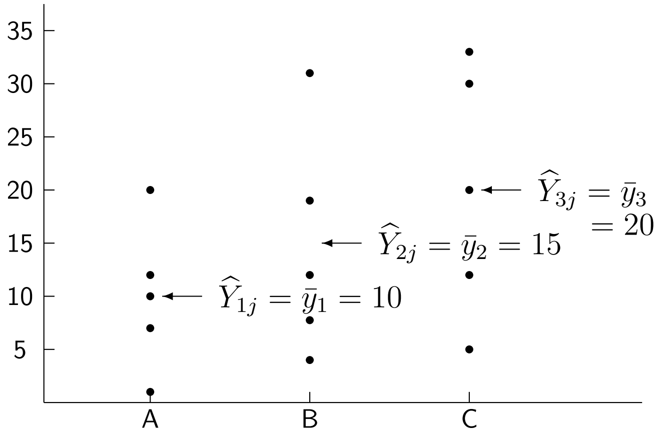
Large MSR: variation within gp
Small $F = \frac{MST}{MSR}$ & accept H_0

Groups		
1	19	5
12	31	33
20	4	20
10	9	12
7	12	30
$\bar{y}_1 = 10$	$\bar{y}_2 = 15$	$\bar{y}_3 = 20$

Reject H_0 of equal means and
Conclude the means are not all equal.



Accept H_0 of equal means and
Conclude the means are all equal.



Data with categorical X and conts. Y: no regression line drawn

15.2 ANOVA model

Let Y_{ij} denote the rv with observed value y_{ij} for $i = 1, 2, .., g$ and $j = 1, 2, ..., n_i$. For each $i = 1, 2, .., g$, we may write

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad j = 1, 2, ..., n_i,$$

where the expected $E(Y_{ij}) = \mu_i$ and the error terms $\epsilon_{ij} = y_{ij} - \mu_i$ which are differences between observed values and their corresponding population means are independent normal rv with

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

for $i = 1, 2, .., g$ and $j = 1, 2, ..., n_i$.

This implies $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ and $Y_{ij} \sim \mathcal{N}(\mu, \sigma^2)$ under $H_0 : \ \mu_1 = \cdots = \mu_g$.

In regression analysis, both the *dependent* Y and *independent* X variables are *continuous*. For example, the data may be (1,2), (3.5,5), (5,9), (6,10) etc.

However in ANOVA, being a *generalization* of the regression analysis, while the *dependent variable* Y is still *continuous*, the X *independent varaible can be categorical* (nominal or ordinal). For example, the data may be (A,2), (A,5), (B,9), (C,10) where $X =$A, B or C gives the label of the group/sample.

15.3 One-way ANOVA test (P.600-607)

The total variations in Y, i.e. $S_{yy} = SST_o$ is measured by the sum of all squared deviation of each point from the overall mean. It is splited into two sources:

1. Variation *between* treatment groups SST: the variation in Y explained by treatment groups,
2. Variation *within* treatment groups SSR: $SST_o - SST$ the variation in Y remain unexplained by treatment groups.

Since $Y_{ij} - \bar{Y}$ can be partitioned as

$$\begin{aligned} Y_{ij} - \bar{Y} &= (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i) \\ \text{Corrected mean} &= \text{Treatment effect} + \text{Residual}, \end{aligned}$$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2}_{SST_o} = \underbrace{\sum_{i=1}^{g} n_i(\bar{Y}_i - \bar{Y})^2}_{SST} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}_{SSR}$$

where

$$\begin{aligned} SST_o &= \text{total corrected sum of squares} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i} Y_{ij}^2 - N\bar{Y}^2, \\ SST &= \text{sum of squares between treatments} = \sum_{i=1}^{g} n_i(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{g} n_i\bar{Y}_i^2 - N\bar{Y}^2, \\ SSR &= \text{sum of squares within treatments} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{g}(n_i - 1)S_i^2. \end{aligned}$$

Then we compare the two sources of variation using F test statistic:

$$f_0 = \frac{MST}{MSR} = \frac{SST/(g-1)}{SSR/(N-g)} = \frac{[\sum_{i=1}^{g} n_i(\bar{Y}_i - \bar{Y})^2]/(g-1)}{[\sum_{i=1}^{g}(n_i-1)S_i^2]/(N-g)}$$

$$= \frac{[\sum_{i=1}^{g}(\frac{\bar{Y}_i-\bar{Y}}{\sigma/\sqrt{n_i}})^2]/(g-1)}{\sum_{i=1}^{g}\frac{(n_i-1)S_i^2}{\sigma^2}/(n_1-1+\cdots+n_g-1)} = \frac{\chi^2_{g-1}/(g-1)}{\chi^2_{N-g}/(N-g)} \sim F_{g-1,N-g}$$

which is the ratio of *mean squared errors* due to *treatment* to that due to residual with rejection region: $f_0 > F_{g-1,N-g;\alpha}$.

The five steps of the test are:

1. **Hypotheses:** $H_0: \mu_1 = ... = \mu_g$ vs
   $H_1:$ Not all means equal.
2. **Test statistic:** $f_0 = \frac{MST}{MSR} = \frac{[\sum_{i=1}^{g} n_i\overline{y}_i^2 - N\overline{y}^2]/(g-1)}{[\sum_{i=1}^{g}(n_i-1)s_i^2]/(N-g)}$.
3. **Assumption:** $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ and Y_{ij} & $Y_{i'j'}$ are independent. Then $f_0 \sim F_{g-1,N-g}$.
4. ***P*-value:** p-value $= \Pr(F_{g-1,N-g} \geq \frac{MST}{MSR})$.
5. **Decision:** Reject H_0 if p-value is less than α.

The calculation for an analysis of variance is usually displayed in an ANOVA table as follows.

One way ANOVA table for completely randomized design

Source	**df**	**SS**	**MS**	**F**
Groups	$g-1$	$SST=\sum_{i=1}^{g} n_i\overline{y}_i^2 - N\overline{y}^2$	$MST=\frac{SST}{g-1}$	$F=\frac{MST}{MSR}$
Residuals	$N-g$	$SSR=\sum_{i=1}^{g}(n_i-1)s_i^2 = SST_o - SST$	$MSR=\frac{SSR}{N-g}=s_p^2$	
Total	$N-1$	$SST_o=\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}^2 - N\overline{y}^2$		

$\overline{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $N=\sum_{i=1}^{g} n_i$ and $\overline{y}=\frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}$.

Remarks:

1. The average $\bar{y}_i$ of observations in the i-th sample and the overall average $\bar{y}$ are

$$\begin{aligned}\bar{y}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},\ \ i=1,\ldots,g;\\ \bar{y} &= \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}.\end{aligned}$$

   When all the group means are equal, i.e. $\bar{y}_1=\cdots=\bar{y}_g=\bar{y}$, $SST=0$ and hence $F=\frac{MST}{MSR}=0$ and we will not reject H_0 of equal means.

2. We predict Y_{ij} by the group mean $\overline{Y}_{i.}$ of treatment group i.

3. The *correction for the mean* is $CM=N(\bar{y})^2=\frac{1}{N}(\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij})^2$.

4. Residual SS (SSR) is used to obtain a pooled estimator of the common variance σ^2.

5. For given SST, the smaller the SSR, the larger the F and hence we will reject H_0. On the other hand, the larger the SST, the stronger is the evidence to indicate a difference between μ_i.

6. The test relies heavily on the assumptions that (a) the equality of variances; (b) the normality of combined observations since under H_0 of equal means, they all come from the same population. We check these assumptions by examining *joint boxplots of the groups* and the *normal qq-plot of the combined residuals.*

7. The ANOVA test is a generalization of the two samples t-test. When $g = 2$, the test statistic for the two sample t-test for $\mu_1 - \mu_2$ assuming equal variance with S_p^2 as its estimate is

$$\begin{aligned}
T^2 &= \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})} = \frac{n_1 n_2 (\bar{Y}_1 - \bar{Y}_2)^2}{\underline{(n_1 + n_2)} S_p^2} \\
&= \frac{\boldsymbol{n_1 n_2 \underline{(n_2 + n_1)}} (\bar{Y}_1 - \bar{Y}_2)^2}{\underline{N^2} S_p^2} \qquad \text{since } N = n_1 + n_2 \\
&= \left\{ \boldsymbol{n_1} \left[\frac{\boldsymbol{n_2} (\bar{Y}_1 - \bar{Y}_2)}{N} \right]^2 + \boldsymbol{n_2} \left[\frac{\boldsymbol{n_1} (\bar{Y}_2 - \bar{Y}_1)}{N} \right]^2 \right\} / S_p^2 \\
&= \left[n_1 \left(\bar{Y}_1 - \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{N} \right)^2 + n_2 \left(\bar{Y}_2 - \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{N} \right)^2 \right] / S_p^2 \\
&= \frac{[n_1(\bar{Y}_1 - \bar{Y})^2 + n_2(\bar{Y}_2 - \bar{Y})^2]/(2-1)}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]/(n_1 + n_2 - 2)} \quad \text{since } \bar{Y} = \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{N} \\
&= \frac{SST/(g-1)}{SSR/(N-g)} = \frac{MST}{MSR} = F_{g-1,N-g} \qquad \text{since } g = 2
\end{aligned}$$

i.e. $F_{1,n-2,\alpha} = t^2_{n-2,\alpha/2}$ and hence more extreme means $F_{1,n-2,\alpha}$ is *more positive* and there is no negative F!

For example,

$t^2_{1,0.025} = 12.706^2 = 161.4 = F_{1,1,0.95}$ $\left(g = 2 \text{ \& } N = 3 \text{ s.t. } g - 1 = 1 \text{ \& } N - g = 1\right)$;
$t^2_{2,0.025} = 4.303^2 = 18.5 = F_{1,2,0.95}$ $\left(g = 2 \text{ \& } N = 4 \text{ s.t. } g - 1 = 1 \text{ \& } N - g = 2\right)$.

16 One-way ANOVA II

16.1 Example

Example: (Teaching techniques) Four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. As a result of dropouts from the experimental groups (due to sickness, transfer, and so on), the number of students varied from group to group.

Do the data shown in the following table present sufficient evidence to indicate a difference in the mean achievement for the four teaching techniques?

Marks

Groups	1	2	3	4
	65	75	59	94
	87	69	78	89
	73	83	67	80
	79	81	62	88
	81	72	83	85
	69	79	76	
		90		

Solution: Let μ_i, $i = 1, 2, 3, 4$, denote the mean mark of the i-th group students (population) and $\{y_{ij}, i = 1, 2, 3, 4, j = 1, .., n_i\}$ denote the mark of jth student in the ith group or population.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_1 : \text{Not all the } \mu_i\text{'s are equal,}$$

The summary of the four treatment groups are

$\bar{y}_i$	75.67	78.43	70.83	87.20	$\bar{y} = 77.667$
s_i^2	66.67	50.62	91.77	26.70	
n_i	6	7	6	5	$N = 24, \quad g = 4$

Also $\sum_{i=1}^{4}\sum_{j=1}^{n_i} y_{ij}^2 = 146736$ and $CM = N\bar{y}^2 = 24(77.667^2) = 144770.7$.

The one-way ANOVA test for the difference in marks across the four teaching techniques is

1. **Hypotheses:** $H_0 : \mu_1 = ... = \mu_4$ vs

   $H_1 :$ Not all means equal.

2. **Test statistic:** $f_0 = \frac{MST}{MSR} = \frac{(\sum_{i=1}^{g} n_i\bar{y}_i^2 - N\bar{y}^2)/(g-1)}{\sum_{i=1}^{g}(n_i-1)s_i^2/(N-g)}$

$$= \frac{762.65/(4-1)}{1202.68/(24-4)} = \frac{254.22}{60.13} = 4.22.$$

$$\begin{aligned}
SST_0 &= \sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}^2 - N\bar{y}^2 = 146736 - 144770.7 = 1965.33 \\
SST &= \sum_{i=1}^{g} n_i(\bar{y}_{i.})^2 - N\bar{y}^2 \\
&= [6(75.67^2) + 7(78.43^2) + 6(70.83^2) + 5(87.20^2)] - 144770.7 \\
&= 762.65 \\
SSR &= SST_0 - SST = 1965.33 - 762.65 = 1202.68 \\
&\overset{or}{=} \sum_{i=1}^{g}(n_i - 1)s_i^2 = 5(66.67) + 6(50.62) + 5(91.77) + 4(26.70)
\end{aligned}$$

3. **Assumption:** $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ and Y_{ij} & $Y_{i'j'}$ are independent. Then $f_0 \sim F_{g-1,N-g}$.

4. ***P*-value:** $\Pr(F_{3,20} \geq 4.22) \in (0.01, 0.025)$

(since $F_{3,20,0.975} = 3.86$, $F_{3,20,0.990} = 4.94$; 0.0182 from R)

5. **Decision:** Since p-value < 0.05, we reject H_0. There are strong evidence in the data against H_0. The marks from groups using different teaching techniques are not all equal.

ANOVA table

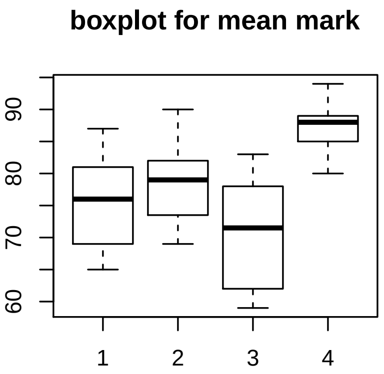
Source	df	SS	MS	F	P-value
Groups	3	762.65	$\frac{762.65}{3} = 254.22$	$\frac{254.22}{60.13} = 4.22$	0.0182
Residuals	20	1202.68	$\frac{1202.68}{20} = 60.13$		
Total	23	1965.33			

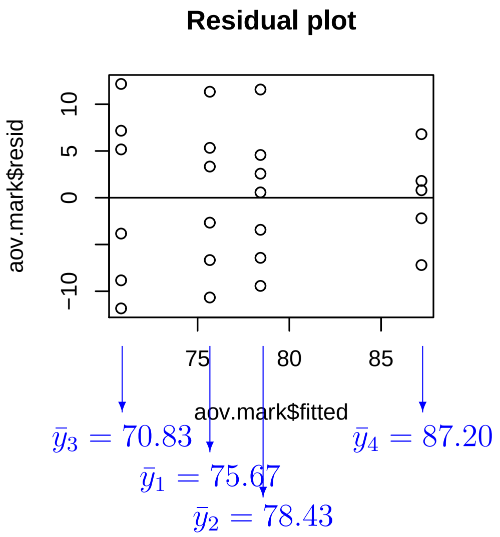
In R,

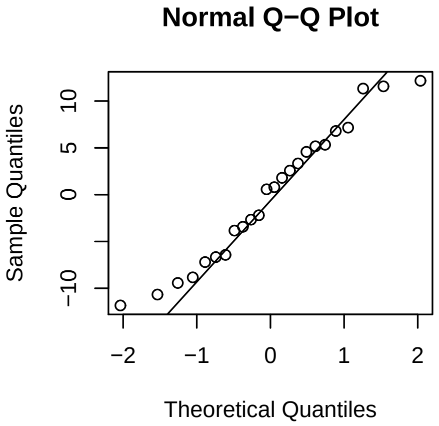
```
> mark=c(65,87,73,79,81,69,75,69,83,81,72,79,90,59,78,67,
      62,83,76,94,89,80,88,85)
> factor=factor(rep(letters[1:4],c(6,7,6,5)))
> factor        # a vector of factor to indicate the groups
 [1] a a a a a a b b b b b b b c c c c c c d d d d d
Levels: a b c d
> aov.mark=aov(mark˜factor)
> summary(aov.mark)
            Df  Sum Sq Mean Sq F value  Pr(>F)
marks.f      3  762.65  254.22  4.2275 0.01812 *
Residuals   20 1202.68   60.13
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> par(mfrow=c(2,2))
> boxplot(mark˜mark.f)  #boxplot for equality of var
> title("boxplot for mean mark")
> plot(aov.mark$fitted,aov.mark$resid)  #residual plot for equality of var
> abline(h=0)
> title("Residual plot")
```

```
> qqnorm(resid(aov.mark))  #check normality
> qqline(resid(aov.mark))
> m=mean(mark)    #for checking
> m
[1] 77.66667
> N=length(mark)
> N
[1] 24
> g=4
> CM=N*m^2
> CM
[1] 144770.7
> meani=tapply(mark,list(mark.f),mean) #mean by groups
> meani
       a        b        c        d
75.66667 78.42857 70.83333 87.20000
> ni=tapply(mark,list(mark.f),length) #length by groups
> ni
a b c d
6 7 6 5
> SST=sum(ni*meani^2)-CM
> SST
[1] 762.6524
> sumsq=sum(mark^2)
> sumsq
[1] 146736
> SSTo=sumsq-CM
> SSTo
[1] 1965.333
> vi=tapply(mark,list(mark.f),var) #var by groups
> vi
       a        b        c        d
66.66667 50.61905 91.76667 26.70000
> SSR=sum((ni-1)*vari)
> SSR
[1] 1202.681
```

```
> f0=(SST/(g-1))/(SSR/(N-g))
> f0
[1] 4.227513
> p.value=1-pf(f0,g-1,N-g)
> p.value
[1] 0.01812315
```


boxplot for mean mark


Residual plot


Normal Q–Q Plot

Note:

1. Since the spreads of each group are quite similar in the *boxplots* and *residual plot*, the *equality of variance* assumption is approximately satisfied.
2. Moreover since the points lie close to the straight line in the *qq-plot*, the *normality* assumption is satisfied.

To show that one-way ANOVA test is a generalization of 2-independent sample t test when $g = 2$, we consider the first two groups in `Marks`:

```
> N.12=n1+n2  #2-indept sample t-test
> SSR.12=v1*(n1-1)+v2*(n2-1)
> MSR.12=SSR.12/(N.12-2)
> t0.12=(m1-m2)/sqrt(MSR.12*(1/n1+1/n2)) #t-test
> pvt.12=2*(1-pt(abs(t0.12),N.12-2))  #2-sided
>
> y.12=c(mark1,mark2)  #one-way ANOVA test
> f.12=factor(rep(letters[1:2],c(n1,n2)))
> summary(aov(y.12~f.12))
            Df Sum Sq Mean Sq F value Pr(>F)
f.12         1   24.6   24.64   0.426  0.528
Residuals   11  637.0   57.91
>
> c(t0.12^2,pvt.12)
[1] 0.4255437  0.5275803
```

Hence $t_0^2 = f_0$ and the 2-sided p-value for t-test and the upper-sided p-value for F-test also agree.

16.2 Transformations for comparisons

What if the two or more groups we compare differ both in location and scale and also show lack of symmetry?

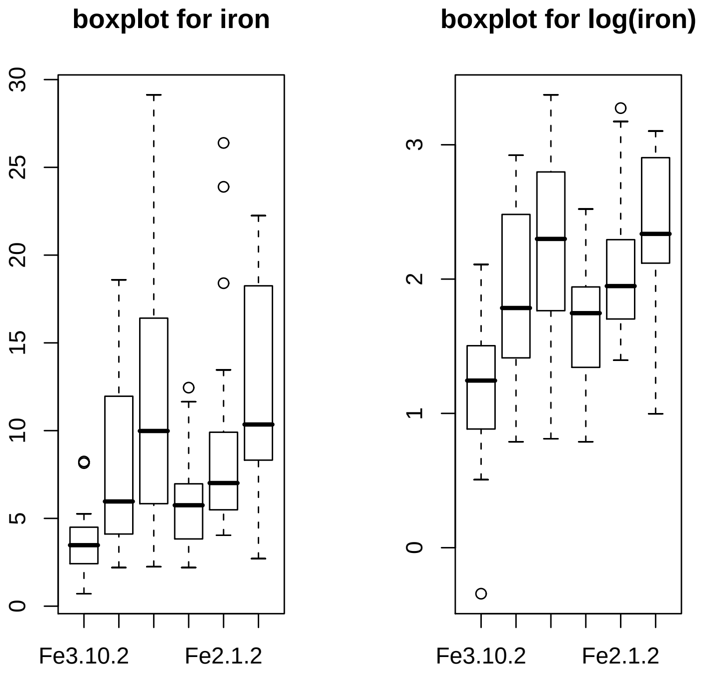
Because the increase in spread with location is often caused by lack of symmetry, it can be remedied by choosing an appropriate transformation which will improve both symmetry within groups and remove differences in spread across groups.

The following data gives retention rates for iron for 6 groups of 18 mice treated with Fe^{2+} and Fe^{3+} at concentrations 10.2, 1.2 and 0.3 millimolar.

View both **iron.csv** and **iron_v.csv** for the matrix and vector formats.

```
> iron=read.csv("data/iron.csv")
> attach(iron)
> iron
 Fe3+.10.2 Fe3+.1.2 Fe3+.0.3 Fe2+.10.2 Fe2+.1.2 Fe2+.0.3
 [1,] 0.71     2.20     2.25      2.20     4.04     2.71
 [2,] 1.66     2.93     3.93      2.69     4.16     5.43
 ...........
[17,] 8.15    18.30    22.82     11.65    23.89    21.60
[18,] 8.24    18.59    29.13     12.45    26.39    22.25
> summary=apply(iron,2,summary)   #summary across col(=2) (note: row=1)
> summary
        Fe3.10.2 Fe3.1.2 Fe3.0.3 Fe2.10.2 Fe2.1.2 Fe2.0.3
Min.       0.710   2.200   2.250    2.200   4.040    2.71
1st Qu.    2.420   4.320   6.103    3.893   5.560    8.50
Median     3.475   5.965   9.980    5.750   7.015   10.35
Mean       3.699   8.204  11.750    5.937   9.632   12.64
3rd Qu.    4.473  11.180  16.000    6.970   9.768   18.16
Max.       8.240  18.590  29.130   12.450  26.390   22.25
> par(mfrow=c(1,2))
> boxplot(iron)  #for matrix format with col
```

```
> title("boxplot for iron")
> boxplot(log(iron))
> title("boxplot for log(iron)")
> ironv=as.matrix(iron)  #for matrix format; change to vector
> ironv=as.vector(ironv)
> ironv
  [1] 0.71 1.66 2.01 2.16 2.42  2.42  2.56  2.60  3.31  3.64  3.74  3.74
 [13] 4.39 4.50 5.07 5.26 8.15  8.24  2.20  2.93  3.08  3.49  4.11  4.95
 [25] 5.16 5.54 5.68 6.25 7.25  7.90  8.85 11.96 15.54 15.89 18.30 18.59
 ...........
> factor=factor(rep(letters[1:6],c(18,18,18,18,18,18))) #for matrix format
> aov.iron=aov(ironv~factor)
> summary(aov.iron)
             Df  Sum Sq Mean Sq F value    Pr(>F)
factor        5 1054.17  210.83  7.3191 6.758e-06 ***
Residuals   102 2938.20   28.81
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> ironlog=log(ironv)
> aov.ironlog=aov(ironlog~factor)
> summary(aov.ironlog)
             Df Sum Sq Mean Sq F value    Pr(>F)
factor        5 18.473   3.695  10.677 2.872e-08 ***
Residuals   102 35.296   0.346
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Remarks

1. As the spreads of the transformed data are closer, the one way ANOVA by using this transformed data is more reliable.
2. As the transformed data is from log(`iron`), the null hypothesis for the one way ANOVA is

$$E(\log Y_1) = E(\log Y_2) = ... = E(\log Y_6).$$

17 Individual comparison (P.608-611)

What if the null hypothesis H_0 of the equality of means is rejected and we would like to know which pairs of means differ?

17.1 Individual comparison

Suppose that groups l and m is believed to have caused the rejection of the null hypothesis:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_g,$$

and we want to test the sub-hypothesis:

$$H_0: \mu_l = \mu_m \quad \text{vs} \quad H_1: \mu_l \neq \mu_m.$$

This sub-hypothesis can be tested by the two samples t-test with test statistic

$$\boxed{t_{gl,gm} = \frac{\bar{x}_{l.} - \bar{x}_{m.}}{\sqrt{MSR}\sqrt{\frac{1}{n_l} + \frac{1}{n_m}}},}$$

where

$\bar{x}_{l.}, \bar{x}_{m.}$ are sample means of groups l and m,
the pooled variance estimate s_p^2 is replaced by MSR and
the d.f. is $N - g$ not $n_l + n_m - 2$.

The corresponding p-value is given by

$$p_{lm} = 2\Pr(t_{N-g} > |t_{lm}|)$$

Also, we obtain a $100(1-\alpha)\%$ confidence interval for $\mu_l - \mu_m$ as

$$\bar{x}_{l.} - \bar{x}_{m.} \pm t_{N-g,\alpha/2}\sqrt{MSR}\sqrt{\frac{1}{n_l} + \frac{1}{n_m}}$$

Note that the MSR using information from all groups will estimate the unknown σ^2 better than the s_p^2 which based on groups l and m only.

Example: (Teaching techniques) We reject the null hypothesis that the mean marks for all groups are equal by using the $ANOVA$-test. From the boxplot, the median marks between group 3 and group 4 are most different and may cause the rejection.

We can check this by using *individual comparison*. We have

$$\bar{x}_{3.} = 70.83, \quad \bar{x}_{4.} = 87.20, \quad MSR = 60.13,$$

$n_3 = 6$ and $n_4 = 5$. Hence a observed value of the two-sample t-statistic is

$$t_{g3,g4} = (70.83 - 87.20)/(\sqrt{60.13}\sqrt{\frac{1}{6} + \frac{1}{5}}) = -3.486,$$

and the corresponding p-value is given by

$$p_{g3,g4} = 2\ \Pr(t_{20} \geq 3.486) = 0.002329398 \quad \text{(from R)}.$$

Conclusion: There are very strong evidence that the mean marks between groups 3 and 4 are different.

17.2 Multiple comparison: the Bonferroni method

What if we want to compare all pairs of means simultaneously?

The idea of the Bonferroni method is that if r null hypotheses are to be tested, a desired overall type I error rate of at most α can be guaranteed by testing each null hypothesis at level $\alpha^* = \alpha/r$.

Let R_i denote the event that the ith null hypothesis is rejected, and let α denote the overall probability of type I error. Then

$$\Pr(R_1 \text{ or } R_2 \text{ or... or } R_r) \le \Pr(R_1) + \Pr(R_2) + ... + \Pr(R_r) \le r\,\alpha^* = r\frac{\alpha}{r} = \alpha.$$

Thus, if each of the g null hypotheses is tested at level α/r, the overall significance level is less than or equal to α.

With g population means $\mu_1, \mu_2, ..\mu_g$, we have $r = \binom{g}{2} = g(g-1)/2$ possible comparisons. To make overall level of significance for the collection of all comparisons considered simultaneously to be less than or equal to α, a test for all l, m is

$$\text{Reject } H_0: \ \mu_l = \mu_m \text{ if } \ p_{l,m} \le \alpha/r,$$

where p_{lm} is the p-value.

Equivalently, we may give a test of the hypothesis:

$$H_0: \ \mu_1 = \mu_2 = ... = \mu_g, \quad \text{vs} \quad H_1: \text{Not all the } \mu_i\text{'s are equal}$$

at the significance level α by

$$\text{Rejecting } H_0 \quad \text{if} \quad \min_{l,m} p_{l,m} \le \alpha/r.$$

Example: (Teaching techniques) The data contain $N = 24$ marks from students under $g = 4$ teaching techniques.

Step 1: Test the equality of means using the ANOVA test: reject H_0 of equal means.
Step 2: Look at a boxplot and a normal `qq`-plot of residual: both assumptions are approximately satisfied.
Step 3: Check differences of pairs of means from the matrix of mean differences.

```
> ni.m=matrix(rep(ni,g),nr=g,byrow=T) #only for unequal group size
> ni.m     #size for col. ; length ni defined before
     [,1] [,2] [,3] [,4]
[1,]    6    7    6    5
[2,]    6    7    6    5
[3,]    6    7    6    5
[4,]    6    7    6    5
> nj.m=t(ni.m)    #transpose, the size for row of the cell
> nj.m
     [,1] [,2] [,3] [,4]
[1,]    6    6    6    6
[2,]    7    7    7    7
[3,]    6    6    6    6
[4,]    5    5    5    5
> meani.m=matrix(rep(meani,g),nr=g,byrow=T) #mean meani defined before
> meani.m     #mean for col.
         [,1]     [,2]     [,3] [,4]
[1,] 75.66667 78.42857 70.83333 87.2
[2,] 75.66667 78.42857 70.83333 87.2
[3,] 75.66667 78.42857 70.83333 87.2
[4,] 75.66667 78.42857 70.83333 87.2
> meanj.m=t(meani.m) #transpose, the mean for row of the cell
> meanj.m
         [,1]     [,2]     [,3]     [,4]
[1,] 75.66667 75.66667 75.66667 75.66667
[2,] 78.42857 78.42857 78.42857 78.42857
```

```
[3,] 70.83333 70.83333 70.83333 70.83333
[4,] 87.20000 87.20000 87.20000 87.20000
> diff=meanj.m-meani.m  #diff. betw. means for the col. & row
> diff   # matrix of pairwise diff.
          [,1]      [,2]      [,3]       [,4]
[1,]  0.000000 -2.761905  4.833333 -11.533333
[2,]  2.761905  0.000000  7.595238  -8.771429
[3,] -4.833333 -7.595238  0.000000 -16.366667
[4,] 11.533333  8.771429 16.366667   0.000000
```

Step 4: Evaluate the observed value of the test statistics.

```
> RMSR=sqrt(sum((ni-1)*vi)/(N-g)) #var vi defined before
> RMSR
[1] 7.754615
> t0=diff/(RMSR*sqrt(1/ni.m+1/nj.m))
> round(t0,3)
       [,1]   [,2]  [,3]   [,4]
[1,]  0.000 -0.640 1.080 -2.456
[2,]  0.640  0.000 1.760 -1.932
[3,] -1.080 -1.760 0.000 -3.485
[4,]  2.456  1.932 3.485  0.000
```

Step 5: Evaluate p-value.

```
> p.value=2*(1-pt(abs(t0),N-g))   #round to 4 dec. place
> round(p.value,4)
       [,1]   [,2]   [,3]   [,4]
[1,] 1.0000 0.5294 0.2930 0.0233
[2,] 0.5294 1.0000 0.0937 0.0677
[3,] 0.2930 0.0937 1.0000 0.0023
[4,] 0.0233 0.0677 0.0023 1.0000
> r=g*(g-1)/2   #no. of pairs  r=4*3/2=6
> c(r,0.05/r)
[1] 6.000000000 0.008333333
> p.value<0.05/r  #result, whether to rej. H0
      [,1]  [,2]  [,3]  [,4]
```

```
[1,] FALSE FALSE FALSE FALSE
[2,] FALSE FALSE FALSE FALSE
[3,] FALSE FALSE FALSE  TRUE
[4,] FALSE FALSE  TRUE FALSE
```

There are differences between means of marks using teaching techniques 3 and 4 only.

Example: (Chemical tested) The data `lab.csv` contain 10 samples of a chemical tested at 7 laboratories is given as follows:

Lab.1	Lab.2	Lab.3	Lab.4	Lab.5	Lab.6	Lab.7
4.13	3.86	4.00	3.88	4.02	4.02	4.00
4.07	3.85	4.02	3.88	3.95	3.86	4.02
4.04	4.08	4.01	3.91	4.02	3.96	4.03
4.07	4.11	4.01	3.95	3.89	3.97	4.04
4.05	4.08	4.04	3.92	3.91	4.00	4.10
4.04	4.01	3.99	3.97	4.01	3.82	3.81
4.02	4.02	4.03	3.92	3.89	3.98	3.91
4.06	4.04	3.97	3.90	3.89	3.99	3.96
4.10	3.97	3.98	3.97	3.99	4.02	4.05
4.04	3.95	3.98	3.90	4.00	3.93	4.06

We want to test whether there are differences in the means of the measurements from the various labs and which pairs of means differ if the differences in the means of the measurements from the various labs are significant.

Solution:

Step 1: Test the equality of means using the ANOVA test.

$$H_0: \ \mu_1 = \mu_2 = ... = \mu_7, \quad \text{vs} \quad H_1: \text{Not all the } \mu_j\text{'s are equal},$$

where μ_j denotes the mean of the measurement in the jth lab.

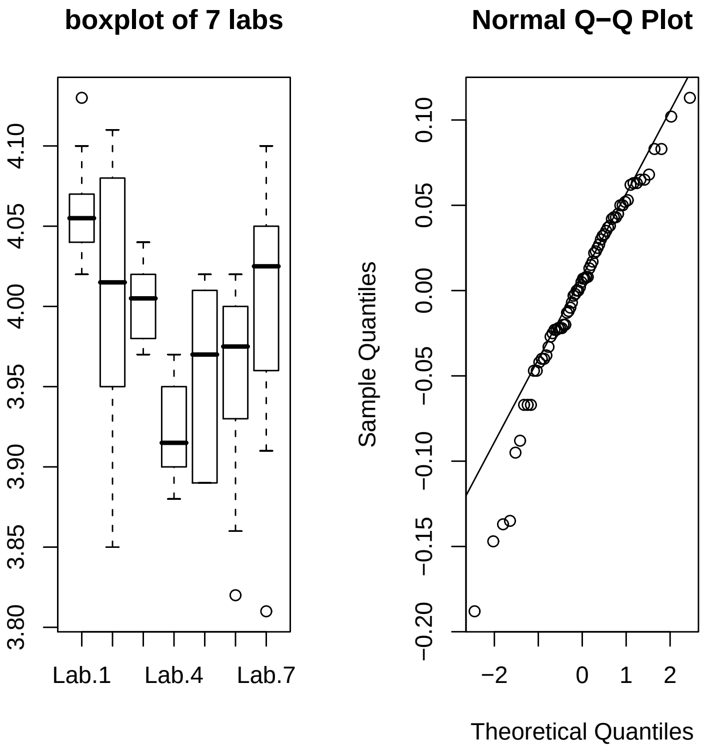
```
> lab=read.csv("lab.csv")
> attach(lab)
> labv=as.matrix(lab)  #for matrix format; to create vector
> labv=as.vector(labv)  #for matrix format; to create vector
> labv
 [1] 4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.10 4.04 3.86 3.85 4.08 4.11
[16] 4.01 4.02 4.04 3.97 3.95 4.00 4.02 4.01 4.01 4.04 3.99 4.03 3.97 3.98
> N=length(labv)
> n=length(lab[,1])  #no. of row is length of each col
> g=length(lab[1,])   #no. of col is length of each row
```

```
> c(N,n,g)
[1] 70 10  7
> lab.f=factor(rep(letters[1:7],c(10,10,10,10,10,10,10))) #for matrix forma
> lab.f
 [1] a a a a a a a a a a b b b b b b b b b b c c c c c c c c c c d d d d d
[39] d d e e e e e e e e e e f f f f f f f f f f g g g g g g g g g g
Levels: a b c d e f g
> aov.lab=aov(labv~lab.f)
> summary(aov.lab)
            Df   Sum Sq  Mean Sq F value    Pr(>F)
lab.f        6 0.124737 0.020790  5.6601 9.453e-05 ***
Residuals   63 0.231400 0.003673
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The p-value is very small. So there are very strong evidence that the means of the measurements from the various labs are significantly different.

Step 2: Look at a boxplot and a normal `qq`-plot of residual.

```
> par(mfrow=c(1,2))
> boxplot(lab)
> title("boxplot of 7 labs")
> qqnorm(resid(aov.lab))
> qqline(resid(aov.lab))
```



These plots indicate that the variances are not similar and that the variables are approximately normal except for the first few points.

Step 3: Check differences of pairs of means from the matrix of mean differences.

```
> g.mean=apply(lab,2,mean)  #col or group mean
> g.mean
Lab.1 Lab.2 Lab.3 Lab.4 Lab.5 Lab.6 Lab.7
4.062 3.997 4.003 3.920 3.957 3.955 3.998
> yi.bar=matrix(rep(g.mean,g),nr=g,byrow=T)  #repeat by row
> yj.bar=t(yi.bar)   #transpose of xi.bar, repeat by col
> diff=yi.bar-yj.bar
> diff
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.000 -0.065 -0.059 -0.142 -0.105 -0.107 -0.064
[2,] 0.065  0.000  0.006 -0.077 -0.040 -0.042  0.001
[3,] 0.059 -0.006  0.000 -0.083 -0.046 -0.048 -0.005
[4,] 0.142  0.077  0.083  0.000  0.037  0.035  0.078
[5,] 0.105  0.040  0.046 -0.037  0.000 -0.002  0.041
[6,] 0.107  0.042  0.048 -0.035  0.002  0.000  0.043
[7,] 0.064 -0.001  0.005 -0.078 -0.041 -0.043  0.000
```

Step 4: Evaluate the observed value of the test statistics.

```
> g.var=apply(lab,2,var)
> round(g.var,5)
  Lab.1   Lab.2   Lab.3   Lab.4   Lab.5   Lab.6   Lab.7
0.00106 0.00805 0.00053 0.00111 0.00327 0.00449 0.00720
> RMSR=sqrt((n-1)*sum(g.var)/(N-g)) #n-1=10-1=9, N-g=70-7=63
> RMSR
[1] 0.06060541
> t0=diff/(RMSR*sqrt(2/n))  #No ni matrix as all ni same
> round(t0,3)  #3 d.p.
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 0.000 -2.398 -2.177 -5.239 -3.874 -3.948 -2.361
[2,] 2.398  0.000  0.221 -2.841 -1.476 -1.550  0.037
[3,] 2.177 -0.221  0.000 -3.062 -1.697 -1.771 -0.184
[4,] 5.239  2.841  3.062  0.000  1.365  1.291  2.878
[5,] 3.874  1.476  1.697 -1.365  0.000 -0.074  1.513
[6,] 3.948  1.550  1.771 -1.291  0.074  0.000  1.587
[7,] 2.361 -0.037  0.184 -2.878 -1.513 -1.587  0.000
```

Note: the test statistic is $t_0 = \dfrac{\bar{y}_i - \bar{y}_j}{RMSR\sqrt{\frac{1}{n}+\frac{1}{n}}} = \dfrac{\bar{y}_i - \bar{y}_j}{RMSR\sqrt{\frac{2}{n}}}$.

Step 5: Evaluate *p*-value.

```
> p.value=2*(1-pt(abs(t0),(N-g)))
> round(p.value,4)  #to 4 dec. place
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 1.0000 0.0195 0.0332 0.0000 0.0003 0.0002 0.0213
[2,] 0.0195 1.0000 0.8258 0.0061 0.1449 0.1262 0.9706
[3,] 0.0332 0.8258 1.0000 0.0032 0.0946 0.0814 0.8546
[4,] 0.0000 0.0061 0.0032 1.0000 0.1771 0.2014 0.0055
[5,] 0.0003 0.1449 0.0946 0.1771 1.0000 0.9412 0.1353
[6,] 0.0002 0.1262 0.0814 0.2014 0.9412 1.0000 0.1175
[7,] 0.0213 0.9706 0.8546 0.0055 0.1353 0.1175 1.0000
```

```
> r=g*(g-1)/2  # no. of pairs  r=7(7-1)/2=21
> round(c(r,0.1/r,0.01/r ),6)
[1] 21.000000  0.004762  0.000476
> p.value<0.1/r  # alpha=0.10
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE
[2,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[3,] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
[4,]  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
[5,]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[6,]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[7,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
> p.value<0.01/r  #alpha=0.01
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,] FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE
[2,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[3,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[4,]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[5,]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[6,]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[7,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

There are differences between means from pairs of labs (1,4), (1,5) and (1,6).

18 Kruskal-Wallis Test (P.677-681)

18.1 The test

If the data from g populations are not well approximated by a normal model or there are obvious differences between the variances of the populations, a *non-parametric* test, the *Kruskal-Wallis* (KW) test may be used.

The *idea* is to rank all the data from all groups together and then apply one-way ANOVA to the ranks rather than to the original data.

Note: KW test is a generalisation of two sample Wilcoxon test in which the observations are replaced by their ranks in the combined sample.

Let r_{ij} be the ranks of $y_{ij}, i = 1, ..., g, j = 1, ..., n_i$. The KW test is

1. **Hypotheses:** $H_0: \ \mu_1 = ... = \mu_g$ vs H_1 : Not all μ_i's are equal.
2. **Test statistic:**

$$k_0 = \frac{SST}{MST_0} = \frac{\sum_{i=1}^{g} n_i(\bar{r}_{i.} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2/(N-1)} = \frac{\sum_{i=1}^{g} n_i \bar{r}_{i.}^2 - N\bar{r}^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N\bar{r}^2/(N-1)}$$

$$\overset{\text{no ties}}{=} \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\,(\bar{r}_i)^2 - 3(N+1)$$

Note: $k_0 = \sum_{i=1}^{g}\left(\frac{\bar{r}_{i.}-\bar{r}}{\hat{\sigma}/\sqrt{n_i}}\right)^2 \sim \chi_{g-1}$, where $\hat{\sigma}^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij}-\bar{r})^2/(N-1)$, as

$$\text{to } F_0 = \frac{MST}{MSR} = \frac{\left[\sum_{i=1}^{g} n_i(\bar{y}_{i.} - \bar{y})^2\right]/(g-1)}{\left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\right]/(N-g)} = \frac{[\sum_{i=1}^{g}(\frac{\bar{y}_i - \bar{y}}{\sigma/\sqrt{n_i}})^2]/(g-1)}{\sum_{i=1}^{g}\frac{(n_i-1)S_i^2}{\sigma^2}/(N-g)},$$

$\bar{r}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}$ is the ith group rank mean,

$\bar{r} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij} = (N+1)/2$, is the overall rank, and

$N = n_1 + n_2 + ... + n_g$ is the overall sample size.

3. **Assumption:** Same distribution of Y_{ij} in each group i. We still have $k_0 \sim \chi^2_{g-1}$ under H_0.
4. ***P*-value:** p-value $= \Pr(\chi^2_{g-1} \geq k_0)$
5. **Decision:** reject H_0 if p-value $< \alpha$.

Remarks

1. If there are ties, the mid-ranks are used as in Wilcoxon tests.
2. In the case of no ties, the proof in Tutorial 7 Q2 shows that

$$\begin{aligned}
\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij} &= \sum_{i=1}^{N} i = N(N+1)/2, \qquad \bar{r} = (N+1)/2, \\
\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 &= \sum_{i=1}^{N} i^2 = \frac{1}{6}N(N+1)(2N+1), \\
\sum_{i=1}^{g}\sum_{j=1}^{n_i} (r_{ij}-\bar{r})^2 &= \sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N\bar{r}^2 = \frac{1}{6}N(N+1)(2N+1) - N\frac{(N+1)^2}{2^2} \\
&= \cdots = N(N+1)(N-1)/12, \\
k_0 &= (N-1)\frac{\sum_{i=1}^{g} n_i\bar{r}_{i.}^2 - N\bar{r}^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N\bar{r}^2} = (N-1)\frac{\sum_{i=1}^{g} n_i\bar{r}_{i.}^2 - \frac{N(N+1)^2}{4}}{\frac{N(N+1)(N-1)}{12}} \\
&\vdots \\
&= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\left(\bar{r}_i\right)^2 - 3(N+1).
\end{aligned}$$

Example: (Chemical tested) The data `lab` contain 10 samples of a chemical tested at 7 laboratories is given as follows:

```
Lab.1 Lab.2 Lab.3 Lab.4 Lab.5 Lab.6 Lab.7
4.13  3.86  4.00  3.88  4.02  4.02  4.00
4.07  3.85  4.02  3.88  3.95  3.86  4.02
4.04  4.08  4.01  3.91  4.02  3.96  4.03
4.07  4.11  4.01  3.95  3.89  3.97  4.04
4.05  4.08  4.04  3.92  3.91  4.00  4.10
4.04  4.01  3.99  3.97  4.01  3.82  3.81
4.02  4.02  4.03  3.92  3.89  3.98  3.91
4.06  4.04  3.97  3.90  3.89  3.99  3.96
4.10  3.97  3.98  3.97  3.99  4.02  4.05
4.04  3.95  3.98  3.90  4.00  3.93  4.06
```

We want to test whether there are differences in the means of the measurements from the various labs. The boxplots show that the variances are quite different across the 7 labs.

Solution: To release the normality and equality-of-variance assumptions, we use the Kruskal-Wallis (KW) test. The summary of the data is

$$\bar{r}_i = 59.85,\ 39.6,\ 39.0,\ 15.5,\ 26.4,\ 26.55,\ 41.6, \qquad \sum_{i=1}^{g} \bar{r}_i^2 = 10043.85,$$

$$\bar{r} = (70+1)/2 = 35.5, \qquad \sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 = 116700.$$

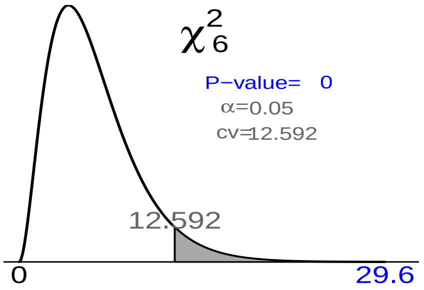
The KW test for the equality of means of measurements from the 7 labs is

1. **Hypotheses:** $H_0: \ \mu_1 = ... = \mu_g$ vs
   H_1 : Not all the μ_i's are equal.

2. **Test statistic:**

$$
\begin{aligned}
k_0 &= \frac{SST}{MST_0} = \frac{\sum_{i=1}^{g} n_i \bar{r}_{i.}^2 - N\bar{r}^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N\bar{r}^2/(N-1)} \\
&= \frac{10(10043.85) - 70(35.5^2)}{[116700 - 70(35.5^2)]/69} = 29.60598
\end{aligned}
$$

3. **Assumption:** Same distribution for Y_{ij} in each group i. We still have $k_0 \sim \chi^2_{g-1}$ under H_0.
4. **P-value:** p-value $= \Pr(\chi^2_6 \geq 29.606) = 0.000047$.
5. **Decision:** Since p-value < 0.05, we reject H_0. There are strong evidence in the data against H_0 that the means of the measurements from the 7 labs are equal.



In R, the function is `kruskal.test(x, groups)` where `x` is a vector of observations and `groups` is a factor of the same length as `x`.

```
> lab=read.csv("lab.csv")
> attach(lab)
> labv=as.matrix(lab) #for matrix format; create vector
> labv=as.vector(labv) #for matrix format; create vector
> labv
 [1] 4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.10 4.04 3.86 3.85 4.08 4.11
[16] 4.01 4.02 4.04 3.97 3.95 4.00 4.02 4.01 4.01 4.04 3.99 4.03 3.97 3.98
[31] 3.88 3.88 3.91 3.95 3.92 3.97 3.92 3.90 3.97 3.90 4.02 3.95 4.02 3.89
[46] 4.01 3.89 3.89 3.99 4.00 4.02 3.86 3.96 3.97 4.00 3.82 3.98 3.99 4.02
```

```
[61] 4.00 4.02 4.03 4.04 4.10 3.81 3.91 3.96 4.05 4.06
> lab.f=factor(rep(letters[1:7],c(10,10,10,10,10,10,10))) #for matrix forma
> lab.f
 [1] a a a a a a a a a a b b b b b b b b b b c c c c c c c c c c
     d d d d d d d d
[39] d d e e e e e e e e e e f f f f f f f f f f g g g g g g g g
     g g
Levels: a b c d e f g
> kruskal.test(labv,lab.f)

        Kruskal-Wallis rank sum test

data:  labv and lab.f
Kruskal-Wallis chi-squared = 29.606, df = 6, p-value = 4.67e-05

> rankv=rank(labv)  #checking only
> rankm=matrix(rankv, nc=7)
> dimnames(rankm)=list(NULL,paste("Lab", 1:7))
> rankm
      Lab 1 Lab 2 Lab 3 Lab 4 Lab 5 Lab 6 Lab 7
 [1,]  70.0   4.5  36.5   6.5  46.5  46.5  36.5
 [2,]  63.5   3.0  46.5   6.5  20.0   4.5  46.5
 [3,]  55.5  65.5  40.5  14.0  46.5  22.5  51.5
 [4,]  63.5  69.0  40.5  20.0   9.0  26.0  55.5
 [5,]  59.5  65.5  55.5  16.5  14.0  36.5  67.5
 [6,]  55.5  40.5  33.0  26.0  40.5   2.0   1.0
 [7,]  46.5  46.5  51.5  16.5   9.0  30.0  14.0
 [8,]  61.5  55.5  26.0  11.5   9.0  33.0  22.5
 [9,]  67.5  26.0  30.0  26.0  33.0  46.5  59.5
[10,]  55.5  20.0  30.0  11.5  36.5  18.0  61.5

> mean.rank =apply(rankm, 2, mean)  #group mean by col
> mean.rank
Lab 1 Lab 2 Lab 3 Lab 4 Lab 5 Lab 6 Lab 7
59.85 39.60 39.00 15.50 26.40 26.55 41.60
> N=length(rankv)
```

```
> g=length(rankm[1,])
> ni=length(rankm[,1])
> c(N,g,ni)
[1] 70  7 10
> sumsq.mrank=sum(mean.rank^2)
> sumsq.rank=sum(rankv^2)
> barmean=mean(rankv)
> c(sumsq.mrank,sumsq.rank,barmean)
[1]  10043.85 116700.00     35.50
> k0=(N-1)*(ni*sumsq.mrank-N*barmean^2)/(sumsq.rank-N*barmean^2)
> p.value=1-pchisq(k0,g-1)
> c(k0,p.value)
[1] 2.960598e+01 4.670102e-05
> k0r=12/(N*(N+1))*ni*sum(mean.rank^2)-3*(N+1)  #only when no ties
> k0r
[1] 29.50757
```

Note that the discrepancy between `k0` and `k0r` is due to the presence of ties.

Summary of test for equality of *means*,

No. of groups	**Parametric** Assume same *Normal* data dist. across groups under H_0	**Non-parametric** Assume same data dist. across groups under H_0
$g = 2$	2-sample t test $t_0 \sim t_{n_x+n_y-2}$	WRS test W or $W_s = \frac{W-E(W)}{\sqrt{Var(W)}} \sim \mathcal{N}(0,1)$
$g \geq 3$	one way ANOVA test $f_0 \sim F_{g-1,N-g}$ $(= t_0^2$ when $g = 2)$	KW test $k_0 \sim \chi^2_{g-1}$ $(= W_s^2$ when $g = 2)$

The proofs of generalization of the 2-sample t test and WRS test are given in this lecture note and Q3, Tutorial 7 respectively.

Note:

1. χ^2 test and F test can be *two-sided* (ratio > 1 or < 1) when applied to the testing of *variances* but ANOVA test and KW test, even though still using F and χ^2 dist., are always *one (upper)-sided* when applied to the testing of several *means* because the test statistics use sum of squares so that some diff. > 0 and others < 0 become all > 0 after squared (upper sided).
2. Ranking should be applied to the *combined sample* for WRS test and KW test assuming same dist. under H_0 but applied to *within each block* for Friedman test (taught later) allowing block effect.
3. ANOVA test assumes *equality of variances* for model assumption but tests for *equality of means*.
4. WRS test and KW test assume same *data* dist. for Y_{ij}, which may not be normal, across groups but use normal approx. to the dist. of *test stat.* W and hence $k_0 = \sum_i W_{s,i}^2 \sim \chi^2_{g-1}$.

To show that the KW test is a generalization of WRS test when $g = 2$, we consider the first two groups of `lab`:

```
> lab=read.csv("lab.csv")
> lab1=lab[,1]
> lab2=lab[,2]
> n=length(lab1)
> rank=rank(c(lab1,lab2))   #WRS test
> N=2*n
> rankA=rank(c(lab1,lab2))[1:n]
> rankA
 [1] 20.0 14.5  9.5 14.5 12.0  9.5  6.5 13.0 18.0  9.5
> rankB=rank(c(lab1,lab2))[(n+1):N]
```

```
> rankB
 [1]  2.0  1.0 16.5 19.0 16.5  5.0  6.5  9.5  4.0  3.0
> w=sum(rankA)
> EW=n*(N+1)/2
> sumsqrank=sum(rank^2)
> g=N*(N+1)^2/4
> varW=n*n*(sumsqrank-g)/(N*(N-1))
> z0=(w-EW)/sqrt(varW)
> p.value=2*(1-pnorm(abs(z0)))  #2-sided
> c(z0^2,p.value)
[1] 2.79301443 0.09467598
>
> lab12v=c(lab1,lab2)
> lab12.f=factor(rep(letters[1:2],c(10,10)))
>
> kruskal.test(lab12v,lab12.f)  #KW test

        Kruskal-Wallis rank sum test

data:  lab12v and lab12.f
Kruskal-Wallis chi-squared = 2.793, df = 1, p-value = 0.09468
```

Hence $z_0^2 = k_0$ and the 2-sided p-value for WSR test and the upper-sided p-value for KW test also agree