Introduction to Geostatistics

9. Two-sample T-test and analysis of variance. Power and Type-II errors.

Edzer Pebesma

edzer.pebesma@uni-muenster.de
Institute for Geoinformatics (**ifgi**)
University of Münster

June 15, 2010

ifgi

How large should a sample be?

Given that a 95% confidence interval, e.g. for μ is obtained by

$$[\bar{X} - t_{df,\alpha}\text{SE}, \bar{X} + t_{df,\alpha}\text{SE}]$$

and given that α is chosen and σ is not under our control, we can only control the width W of the interval by manipulating n:

$$W = 2t_{df,\alpha}\text{SE} = 2t_{df,\alpha}s/\sqrt{n}$$

$$n = (\frac{2t_{df,\alpha}s}{W})^2$$

How about controlling type II errors?

ifgi

Type I and Type II errors

Of course we take a risk to wrongly rejecting a true H_0, of α. There's however also a risk that we wrongly *not* reject a false H_0, which is called β.

Test result	Truth	
	H_0 true	H_0 false
Reject H_0	Type I error, α	OK, (1-β)
Do not reject H_0	OK (1-α)	Type II error, β

Next 2 slides from: Wonnacott & Wonnacott, Introductory statistics.

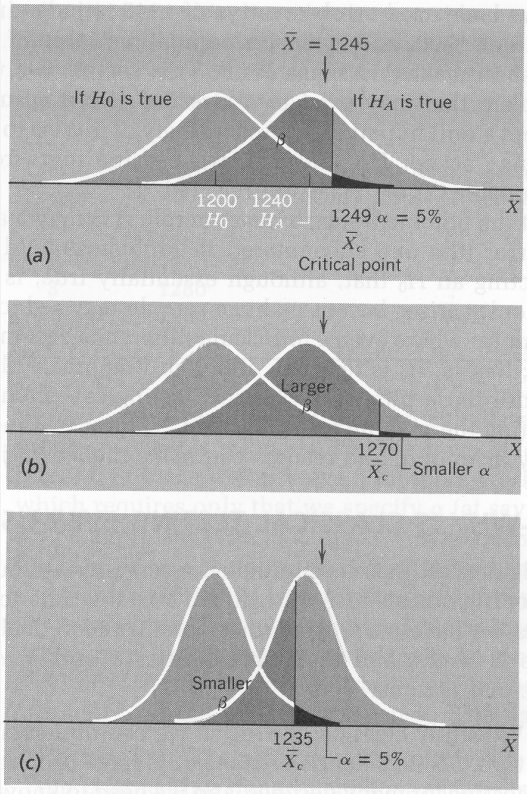
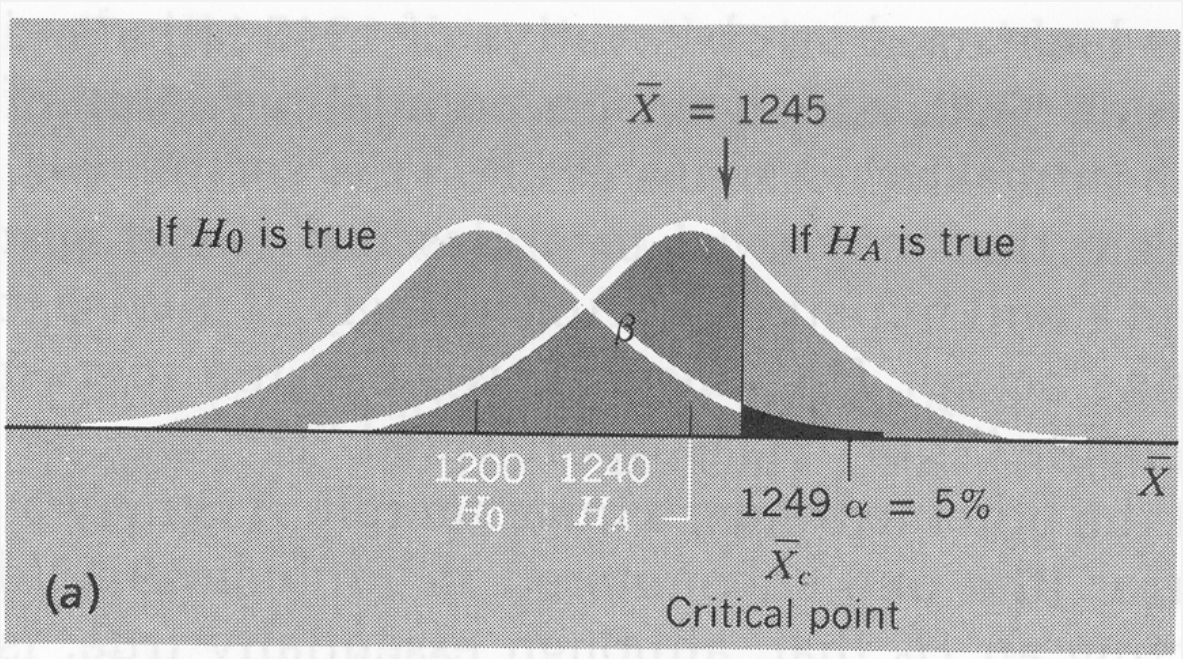


FIGURE 9-7
(a) Hypothesis test of Figure 9-6 showing α and β. (b) How a reduction in α increases β, other things being equal. (c) How an increase in sample size allows one error probability (β) to be reduced, without increasing the other (α).



ifgi

How to compute the power function?

- Given that H_0 is not true, then what is true? Probabilities cannot be computed without assumptions about the population.
- Given a fixed H_A, we can compute power as in the figure in the previous slide.
- For all possible H_A's, we obtain the *power function*.
- What determines the power?
  - The difference between the H_0 and H_A means (delta)
  - The width of the curves (SE $= \sigma/\sqrt{n}$)
  - α
  - where is α? – one-sided or two-sided
  - what is n? how is SE computed? – type of test: one-sample, two-sample, paired

ifgi

Power computation using `power.t.test`

Description: Compute power of test, or determine parameters to obtain target power.
Details: Exactly one of the parameters `n`, `delta`, `power`, `sd`, and `sig.level` must be passed as NULL, and that parameter is determined from the others. Notice that the last two have non-NULL defaults so NULL must be explicitly passed if you want to compute them.

Compute sample size

```
> power.t.test(n = NULL, delta = 1, sd = 1, sig.level = 0.05,
+     power = 0.9, type = "two.sample", alternative = "two.sided")

     Two-sample t test power calculation

              n = 22.02110
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

Compute delta (H_A)

```
> power.t.test(n = 20, delta = NULL, sd = 1, sig.level = 0.05,
+     power = 0.9, type = "two.sample", alternative = "two.sided")

     Two-sample t test power calculation

              n = 20
          delta = 1.051970
             sd = 1
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

Compute power

```
> power.t.test(n = 20, delta = 1, sd = 1, sig.level = 0.05,
+     power = NULL, type = "two.sample", alternative = "two.sided")

     Two-sample t test power calculation

              n = 20
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8689528
    alternative = two.sided

 NOTE: n is number in *each* group
```

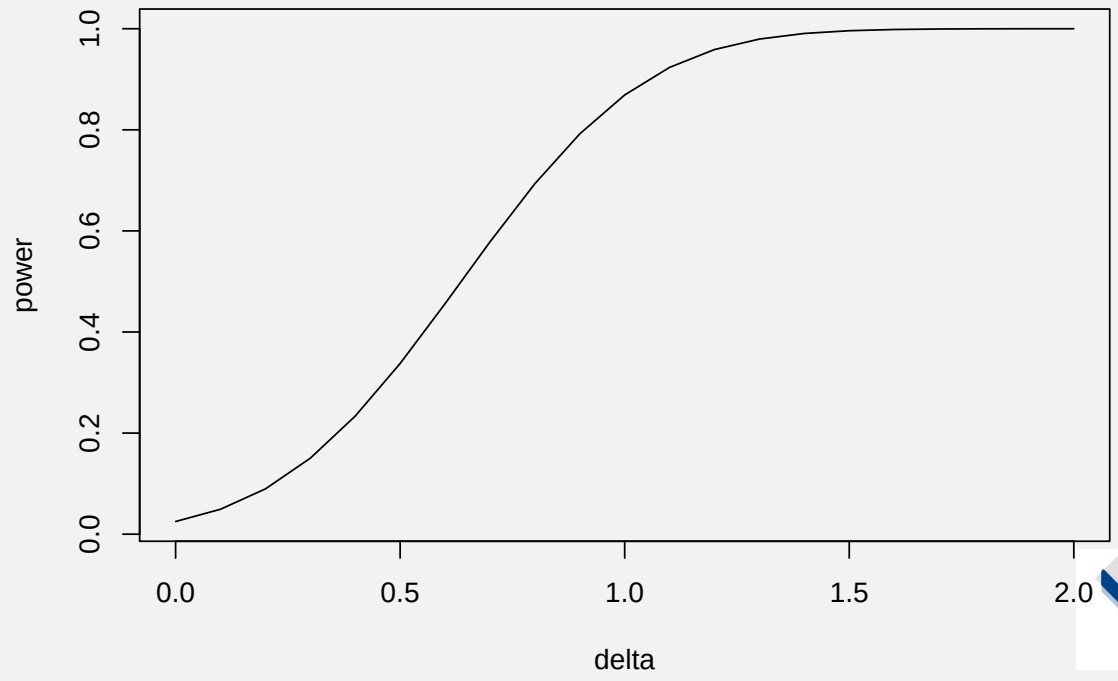
Compute significance level

```
> power.t.test(n = 20, delta = 1, sd = 1, sig.level = NULL,
+     power = 0.9, type = "two.sample", alternative = "two.sided")

     Two-sample t test power calculation

              n = 20
          delta = 1
             sd = 1
      sig.level = 0.07004584
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

(Note that this is of little operational use; computing sd is of even less operational use)

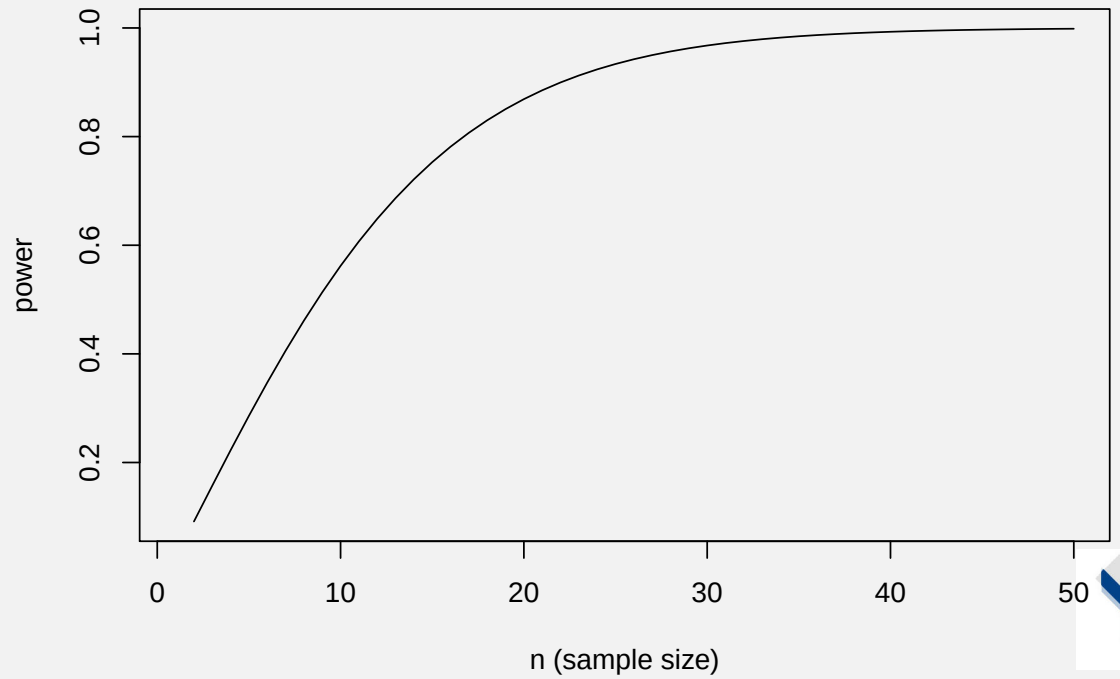
Compute power function vs. delta, n = 20, s = 1

```
> plot((0:20)/10, power.t.test(power = NULL, delta = (0:20)/10,
+     n = 20)$power, type = "l", xlab = "delta", ylab = "power")
```



Compute power function vs. n; delta = 1

```
> plot(1:50, power.t.test(delta = 1, n = 1:50)$power, type = "l",
+     xlab = "n (sample size)", ylab = "power")
```



ifgi

The power concept beyond n

In a testing framework, increasing *n* will make every small difference in means significant, as small differences will be noted (with large power). This does not mean that the difference found is relevant.

Suppose we're studying the effect of a medication type on health, or a herbicide type on plant disease. Two large samples (with and without treatment) confirmed (showed significantly) that in the group without treatment there was 45% succes, less than in the group with treatment with 47% success.

That's OK, but should we now collectively apply the treatment? Do the effects compensate for the costs and side effects?

Significance is something else as relevance

ifgi

The power concept beyond n

Taking a larger sample always increases power. Can we do something else to increase power? Yes: choose a more appropriate analysis. Recall the paired data of lecture 7:

obj	t_1	t_2
1	13.5	12.7
2	15.3	15.1
3	7.5	6.6
4	10.3	8.5
5	8.7	8.0

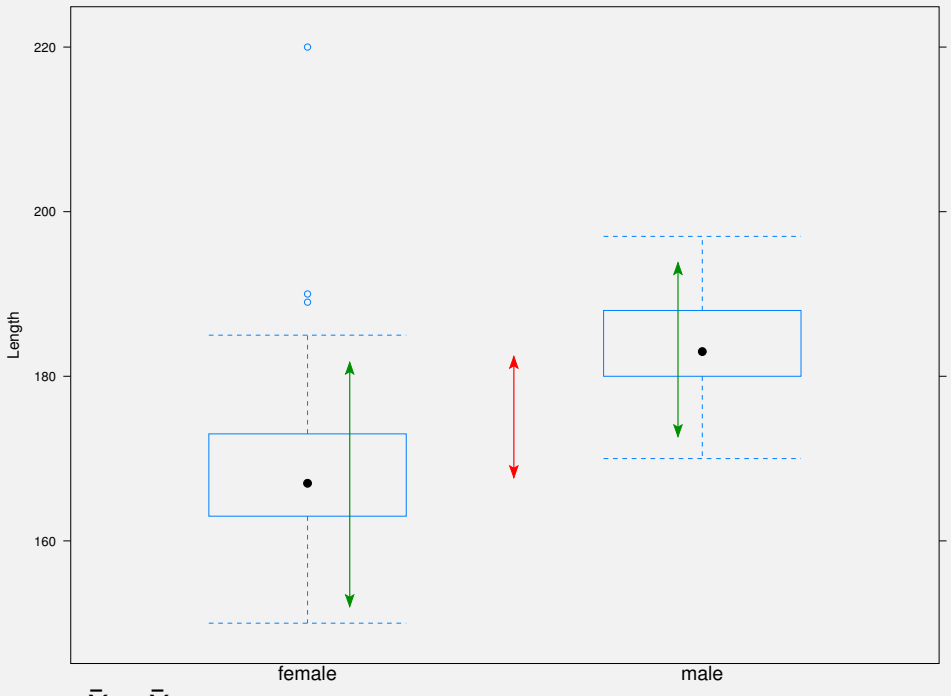
```
> x1 = c(13.5, 15.3, 7.5, 10.3, 8.7)
> x2 = c(12.7, 15.1, 6.6, 8.5, 8)
> x1 - x2

[1] 0.8 0.2 0.9 1.8 0.7
```


```
> t.test(x1, x2, var.equal = TRUE)

        Two Sample t-test

data:  x1 and x2
t = 0.4066, df = 8, p-value = 0.695
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.111314  5.871314
sample estimates:
mean of x mean of y
    11.06     10.18

> t.test(x1, x2, paired = TRUE)

        Paired t-test

data:  x1 and x2
t = 3.3896, df = 4, p-value = 0.02754
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1591929 1.6008071
sample estimates:
mean of the differences
                   0.88
```

```
> power.t.test(delta = 0.88, n = 5, sd = sqrt((var(x1) +
+     var(x2))/2))

     Two-sample t test power calculation

              n = 5
          delta = 0.88
             sd = 3.422353
      sig.level = 0.05
          power = 0.0548756
    alternative = two.sided

 NOTE: n is number in *each* group

> power.t.test(delta = 0.88, n = 5, sd = sd(x1 - x2), type = "paired")

     Paired t test power calculation

              n = 5
          delta = 0.88
             sd = 0.580517
      sig.level = 0.05
          power = 0.7192318
    alternative = two.sided

 NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

ifgi

Two-sample T-test and analysis of variance.



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s/\sqrt{n}}$$

ifgi

Generalizing 2 to p groups: from t to F

- let $n_1 = n_2 = n$ and $N = n_1 + n_2$, and assume $\sigma_1 = \sigma_2$
- $t = \frac{\bar{X}_1 - \bar{X}_2}{s/\sqrt{n}} = \frac{(\bar{X}_1 - \bar{X}_2)\sqrt{n}}{s}$
- $t^2 = \frac{(\bar{X}_1 - \bar{X}_2)^2 n}{s^2} = \frac{n\text{Var}(\bar{X}_i)}{s^2}$
- $\frac{n\text{Var}(\bar{X}_i)}{s^2}$, with s^2 the pooled (averaged, joined) within-group variance
- numerator: variance, as obtained from variability between groups (group means)
- denominator: variance, as obtained from variability within groups (ignores differences between groups)
- Under the hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_p$,

$$F = \frac{n\text{Var}(\bar{X}_i)}{s^2}$$

follows the F distribution with $p - 1$ (numerator) and $N - p$ (denominator) degrees of freedom.

This idea generalizes the two-sample t-test, testing $H_0 : \mu_1 = \mu_2$ to the F-test, testing $H_0 : \mu_1 = \mu_2 = ... = \mu_p$.

ifgi

Why not use many t-tests?

- Suppose we have three groups, and we can reject $H_0 : \mu_1 = \mu_2$, we can reject $H_0 : \mu_2 = \mu_3$, but cannot reject $H_0 : \mu 1 = \mu_3$.
  This will be clumsy to explain.
  When hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ is rejected, we can continue work under the hypothesis "the group means are not identical".
- Suppose we have many (10) groups with few observations (3) each. Pairwise testing has very little power (df = 4), whereas joint testing with ANOVA has (df = 20).

ifgi

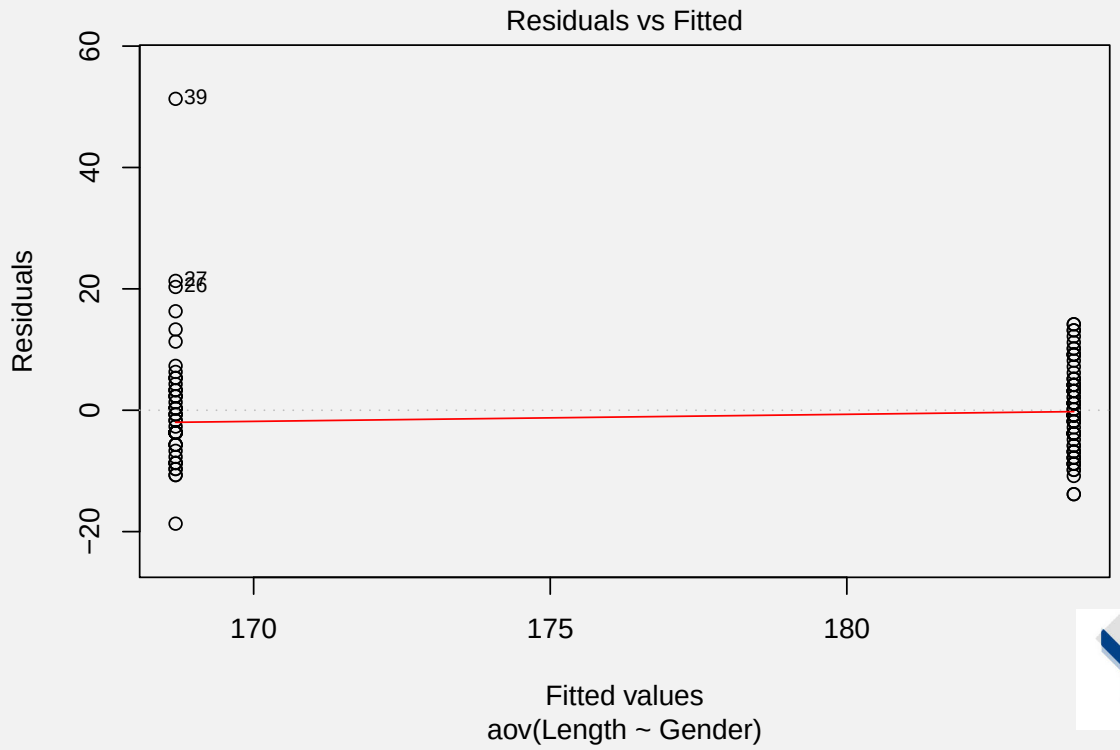
How to read ANOVA tables?

```
> summary(aov(Length ~ Gender))
             Df Sum Sq Mean Sq F value    Pr(>F)
Gender        1 4017.1  4017.1  45.466 1.845e-09 ***
Residuals    84 7421.8    88.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`Gender`: effect, explanatory variable, grouping variable, between groups
`Residuals`: error, within-groups, unexplained variability
`Df`: degrees of freedom for that row
`Sum Sq`: sum of squares, between or within
`Mean Sq`: mean squares: `Sum Sq` divided by `Df`
`F value`: `Mean Sq` effect divided by `Mean Sq` Residuals
`Pr(>F)`: significance level, p-value

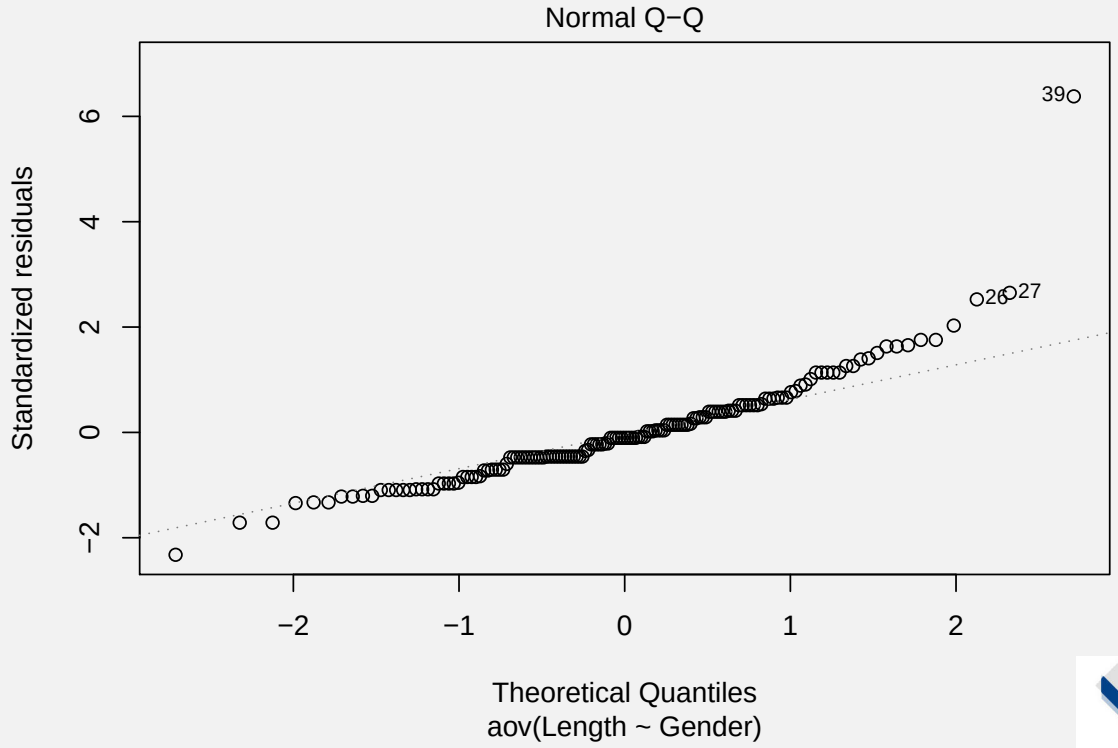
ifgi

```
> plot(aov(Length ~ Gender), which = 1)
> Length[31]
[1] 162
```

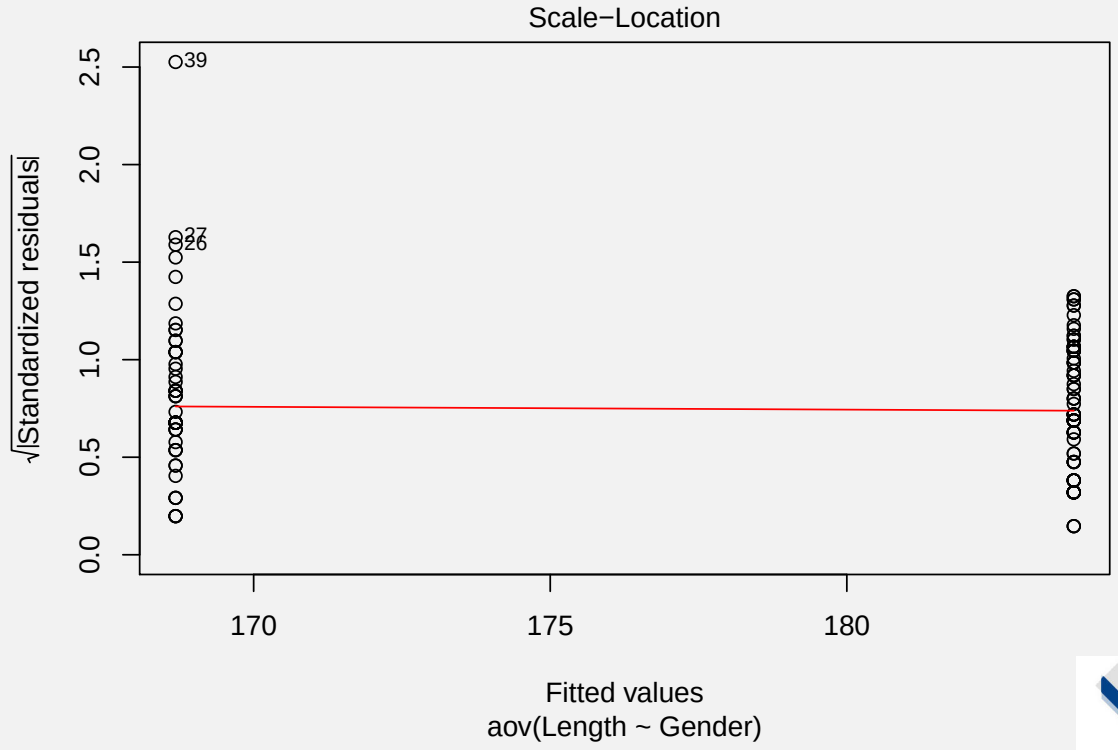


ifgi

```
> plot(aov(Length ~ Gender), which = 2)
```




```
> plot(aov(Length ~ Gender), which = 3)
```



Two-way ANOVA; setting up data

The data can also be organized like this:

```
> x = data.frame(resp = c(x1, x2), time = rep(c("t1", "t2"),
+     each = 5), obj = rep(letters[1:5], 2))
> x
```

```
   resp time obj
1  13.5   t1   a
2  15.3   t1   b
3   7.5   t1   c
4  10.3   t1   d
5   8.7   t1   e
6  12.7   t2   a
7  15.1   t2   b
8   6.6   t2   c
9   8.5   t2   d
10  8.0   t2   e
```

Two-way ANOVA

One-way ANOVA:

```
> summary(aov(resp ~ time, x))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
time         1  1.936   1.936  0.1653  0.695
Residuals    8 93.700  11.713
```

Two-way ANOVA:

```
> summary(aov(resp ~ time + obj, x))
```

```
            Df Sum Sq Mean Sq F value    Pr(>F)
time         1  1.936  1.9360  11.490 0.0275393 *
obj          4 93.026 23.2565 138.021 0.0001545 ***
Residuals    4  0.674  0.1685
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Paired t-test vs two-way ANOVA

Paired t-test:

```
> t.test(x1, x2, paired = TRUE)

        Paired t-test

data:  x1 and x2
t = 3.3896, df = 4, p-value = 0.02754
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1591929 1.6008071
sample estimates:
mean of the differences
                   0.88
```

Two-way ANOVA:

```
> summary(aov(resp ~ time + obj, x))

            Df Sum Sq Mean Sq F value    Pr(>F)
time         1  1.936  1.9360  11.490 0.0275393 *
obj          4 93.026 23.2565 138.021 0.0001545 ***
Residuals    4  0.674  0.1685
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ifgi

Paired t-test vs. two-way ANOVA

Note: *p*-values are identical. Anova generalizes paired t-tests in the sense that e.g. time can have more than 2 levels (but is considered categorical).
Further extensions: three-way, more-way anova; interactions.
Now introduce the `meuse` data set

ifgi