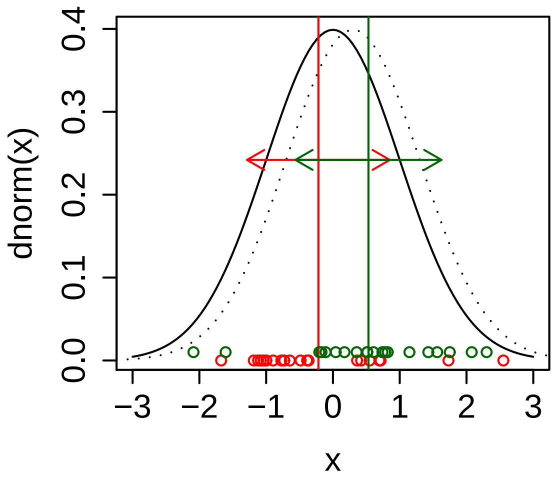
Use of SAS
December, 2010

3. t-test and One-way ANOVA

Comparing two samples

Two groups,

$x_{11}, \ldots, x_{1n_1} \qquad x_{21}, \ldots, x_{2n_2}$



$N(\mu_1, \sigma_1^2)$

$N(\mu_2, \sigma_2^2)$

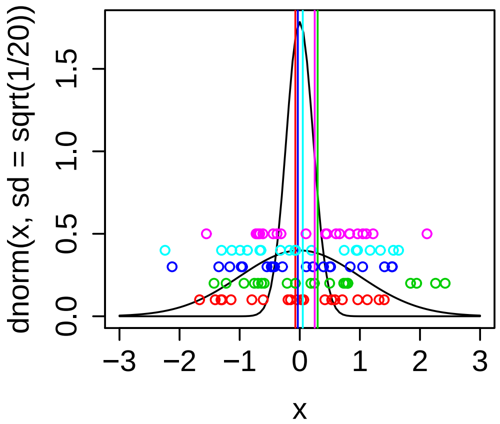
$(\bar{x}_1, s_1^2)$

$(\bar{x}_2, s_2^2)$

Significant difference between $\bar{x}_1$ and $\bar{x}_2$?

Null hypothesis $H_0 : \mu_1 = \mu_2$

Two-sample t-test



$$\mathrm{SEM} = s/\sqrt{n}$$

Standard error of mean

$$\mathrm{SEDM} = \sqrt{\mathrm{SEM}_1^2 + \mathrm{SEM}_2^2}$$

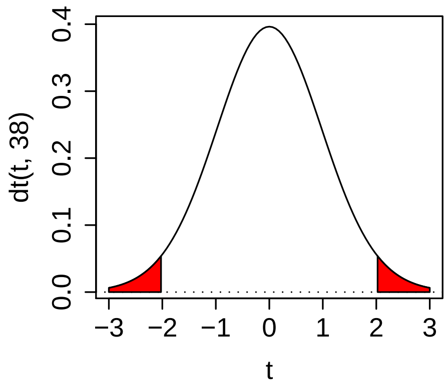
Standard error of difference of means

$$\boxed{t = \frac{\bar{x}_2 - \bar{x}_1}{\mathrm{SEDM}}}$$

test-statistic t measures disagreement between data and H_0

The p-value

$$\boxed{t = \frac{\bar{x}_2 - \bar{x}_1}{\text{SEDM}}}$$



t: measures disagreement between data and H_0

If H_0 is true: distribution of t is symmetric around 0

p: the prob. of having observed a more extreme t-value

if $p < 5\%$: H_0 is rejected

Two tests: same or different variances?

Assume $\sigma_1^2 = \sigma_2^2$ before testing $\mu_1 = \mu_2$?

Same variance:

- Natural under null hypothesis (same distributions)
- Nice theory.

Separate variances:

- Looks specifically for difference in means
- Approximative theory.

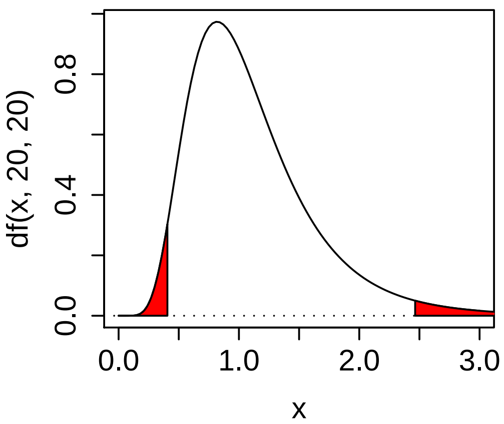
Test for same variance

Test statistic

$$F = s_1^2 / s_2^2$$

F-distribution with (f_1, f_2) degrees of freedom

$f_1 = n_1 - 1 \qquad f_2 = n_2 - 1$



Note: 2-sided test

t-test in SAS

Example from Anders Juul

IDEA: Compare men and women with respect to $\sqrt{\mathsf{SIGF1}}$ for 20–30 year olds.

1. Open dataset: SET statement;
2. Compute $\mathsf{SSIGF1} = \mathsf{SQRT(SIGF1)}$
3. Start PROC TTEST
4. Use WHERE to select subgroup
5. Specify dependent variable
6. — and classification
7. Do not forget RUN

Code for the t-test

```
data juul;
   set sasuser.juul;
   ssigf1=sqrt(sigf1);
run;
proc ttest data=juul;
   where age > 20 and age < 30;
   var ssigf1;
   class sexnr;
run;
```

Output

The TTEST Procedure

Statistics

Variable	sexnr	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
ssigf1	1	23	15.789	16.517	17.245	1.3021	1.6836	2.3829	0.3511
ssigf1	2	18	15.798	16.824	17.85	1.5481	2.063	3.0928	0.4863
ssigf1	Diff (1-2)		-1.49	-0.307	0.8763	1.5225	1.8586	2.3864	0.5849

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
ssigf1	Pooled	Equal	39	-0.52	0.6030
ssigf1	Satterthwaite	Unequal	32.5	-0.51	0.6125

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
ssigf1	Folded F	17	22	1.50	0.3666

Exercise: T-test

Consider again the Juul data with variables

- Age (years)
- Height (cm)
- Menarche (No/Yes: 1/2)
- Sex (M/F: 1/2)
- Serum IGF1, growth hormone (μg/ml)
- Tanner stage (1–5)
- Testis volume (ml)
- Weight (kg)

Here the main aim is to compare the IGF1-level in boys and girls *above the age of 5 years*.

1. For each Tanner-stage, test if the IGF1-level is the same in boys and girls. The distribution of IGF1 seems to be skew, but $\sqrt{SIGF1}$ can be assumed to follow a normal distribution.

One-way ANOVA

Comparing more than two groups

$\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k \qquad s_1, s_2, \ldots, s_k$

Joint test for any differences between the groups.

Why not just pairwise t-tests?

MASS SIGNIFICANCE
LOSS OF OVERVIEW

The fewer tests, the better.

Notation and Models

x_{ij} observation no. j in group no. i

(i.e., x_{35} the 5th observation in group 3)

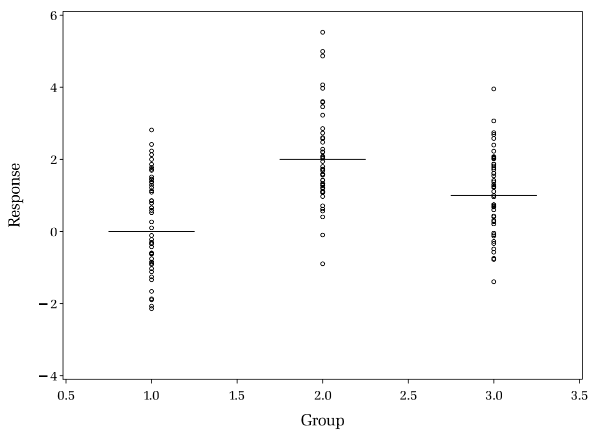
The model

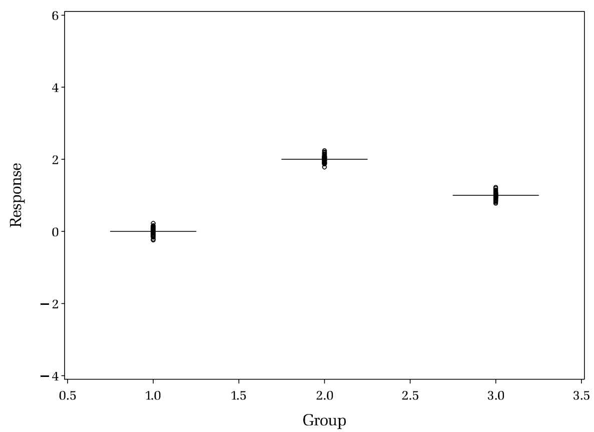
$$X_{ij} = \mu_i + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

The hypothesis of no differences between groups

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

Variation within and between groups





Main idea:

If the variation between group means is large compared to the variation within groups, it is a sign that the hypothesis is wrong.

Sums of squares

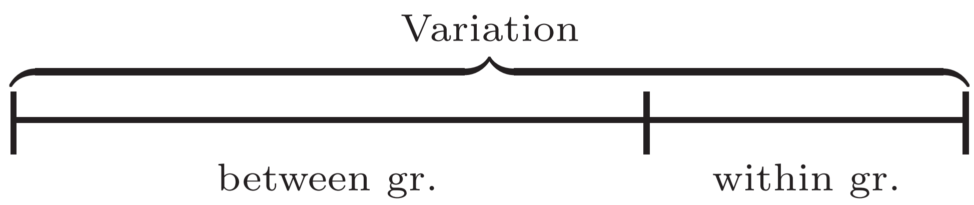
Variation (**W**ithin) groups: $\mathrm{SSD}_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$
$\bar{x}_i$ mean for group i

Variation (**B**etween) groups: $\mathrm{SSD}_B = \sum_i \sum_j (\bar{x}_i - \bar{x}_.)^2$
$\bar{x}_.$ total (grand) mean

Can be mathematically proven that

$$\mathrm{SSD}_B + \mathrm{SSD}_W = \mathrm{SSD}_{\mathrm{total}} = \sum_i \sum_j (x_{ij} - \bar{x}_.)^2$$

The model (grouping) *explains* part of the variation

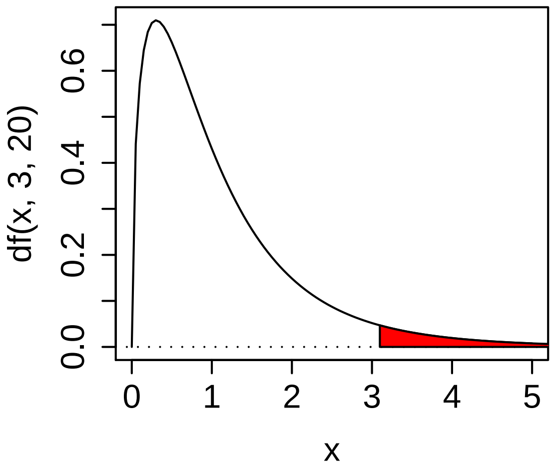


F-test for identical group means

We reject the hypothesis if the variation between groups is large compared to the variation within groups. Consider

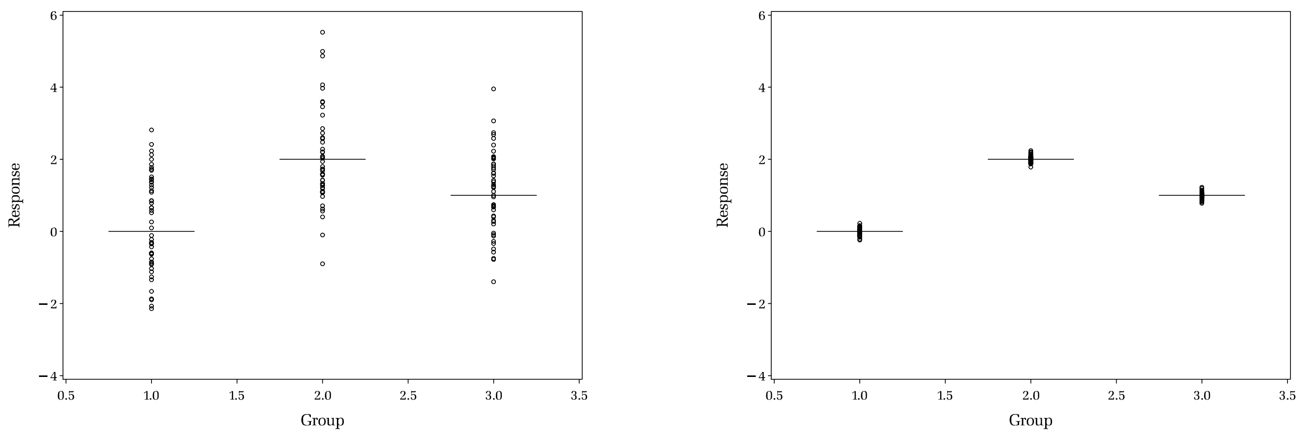
$$F = [\mathrm{SSD}_B/(k-1)]/[\mathrm{SSD}_W/(N-k)]$$

If group differences are coincidental then F follows an F-distribution:



If F too large: Reject the hypothesis that the groups are identical.

Testing for identical group means



high variation within grp.

F is small

H_0 is *not* rejected

small variation within grp.

F is large

H_0 is rejected

One-way ANOVA in SAS

IDEA: Compare boys in different Tanner stage with respect to their $\sqrt{\mathsf{SIGF1}}$

1. This time generate a new data set, `juulboys`
2. Select: $\mathsf{SEXNR} = 1$, $\mathsf{AGE} < 20$
3. Use GLM
4. MODEL statement: What is described by what?
5. Remember to say that `tanner` is a grouping (CLASS)

Code for ANOVA

```
data juulboys;
   set sasuser.juul;
   ssigf1 = sqrt(sigf1);
   if sexnr = 1 and 0 < age < 20;
run;
proc glm data=juulboys;
   class tanner;
   model ssigf1 = tanner / solution;
run;
```

(PROC ANOVA can also be used)

Output

```
                          The GLM Procedure

                      Class Level Information

                  Class         Levels    Values

                  tanner             5    1 2 3 4 5


                    Number of observations    546

NOTE: Due to missing values, only 400 observations can be used in this analysis.

                          The GLM Procedure

Dependent Variable: ssigf1

                                       Sum of
      Source                  DF      Squares     Mean Square    F Value    Pr > F

      Model                    4   6054.50950      1513.62738     147.14    <.0001

      Error                  395   4063.35801        10.28698

      Corrected Total        399  10117.86751
```

R-Square	Coeff Var	Root MSE	ssigf1 Mean
0.598398	18.35978	3.207333	17.46934

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tanner	4	6054.509502	1513.627376	147.14	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tanner	4	6054.509502	1513.627376	147.14	<.0001

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		21.49657843 B	0.29908531	71.87	<.0001
tanner	1	-7.93544551 B	0.37819314	-20.98	<.0001
tanner	2	-3.24333496 B	0.60013505	-5.40	<.0001
tanner	3	-0.19150204 B	0.73260639	-0.26	0.7939
tanner	4	1.26129188 B	0.64103059	1.97	0.0498
tanner	5	0.00000000 B	.	.	.

Nonparametric tests

Mann-Whitney test (alias Wilcoxon)

Kruskal-Wallis test

Idea: “t-test” or “ANOVA” on *ranks*

$$\underset{\text{Data}}{\boxed{\begin{array}{c} 3, 9, 17, 50 \\ 4, 7, 8.7, 200 \end{array}}} \longrightarrow \underset{\text{Rank}}{\boxed{\begin{array}{c} 1, 5, 6, 7 \\ 2, 3, 4, 8 \end{array}}}$$

Distribution is (in principle) known under null hypothesis. Does not depend on data following a Normal distribution.

Other “scores” than ranks can also be used

Nonparametric tests in SAS

- ```
  proc npar1way wilcoxon data=sasuser.juul;
      where sexnr = 1 and 0 < age < 20;
      var sigf1;
      class tanner;
  run;
  ```
- (Mann-Whitney is obtained if there are only two groups to compare)
- It is the `wilcoxon` option that select rank scores and thus the Kruskal-Wallis/Mann-Whitney test, see manual for alternatives.
  ```

Output

```
                    The NPAR1WAY Procedure

         Wilcoxon Scores (Rank Sums) for Variable sigf1
                  Classified by Variable tanner

                         Sum of      Expected       Std Dev          Mean
tanner          N        Scores      Under H0      Under H0         Score
1             192      20758.00      38496.00    1155.20917    108.114583
2              38       8222.00       7619.00     677.99178    216.368421
3              23       6569.50       4611.50     538.28582    285.630435
4              32      10387.00       6416.00     627.30283    324.593750
5             115      34263.50      23057.50    1046.52522    297.943478

                Average scores were used for ties.


                      Kruskal-Wallis Test

                   Chi-Square         254.3465
                   DF                        4
                   Pr > Chi-Square      <.0001
```