One Factor ANOVA

What is Experimental Design?

- A designed experiment is a test in which purposeful changes are made to the input variables (x) so that we may observe and identify the reasons for change in the output response (y).
- Objectives:
  - Which variables are most influential on the response y?
  - Where to set the influential x's so y is near a desired value?
  - Where to set the influential x's so variability in y is small?

Principals of design

- Replication: repetition of the basic experiment
  - Allows the estimation experimental error
  - Obtain a more precise estimate of the effect of a factor
- Randomization: both the allocation of the experimental material and the order in which the individual runs are performed are randomly determined
- Blocking: a block is a homogeneous portion of the experimental material
  - Used to increase the precision of an experiment

Business Example

- A manufacturer is interested in **maximizing** the tensile strength of a new synthetic fiber that will be used to make cloth for men's shirt.
- From previous tests, the manufacturer knows:
  - the strength is affected by the **percentage of cotton** in the fiber
  - The **range** of the percentage is 10% to 40%
- Experiment with a Single Factor

Example – continued

- Maximizing the tensile strength of a new synthetic fiber used to make shirts*
- Strength is affected by the % of cotton in the fiber
  - Test five levels of cotton percentage
    - 15%, 20%, 25%, 30%, 35%
  - Test five specimens at each cotton %
  - So, a = 5 levels of the factor, and n = 5 replicates
  - All 25 runs are made in random order

*DCM pp 50

Example – continued

Experimental Runs

- This randomized test sequence, known as a Completely Randomized Design (CRD), is necessary to prevent the effects of unknown nuisance variables from contaminating the results

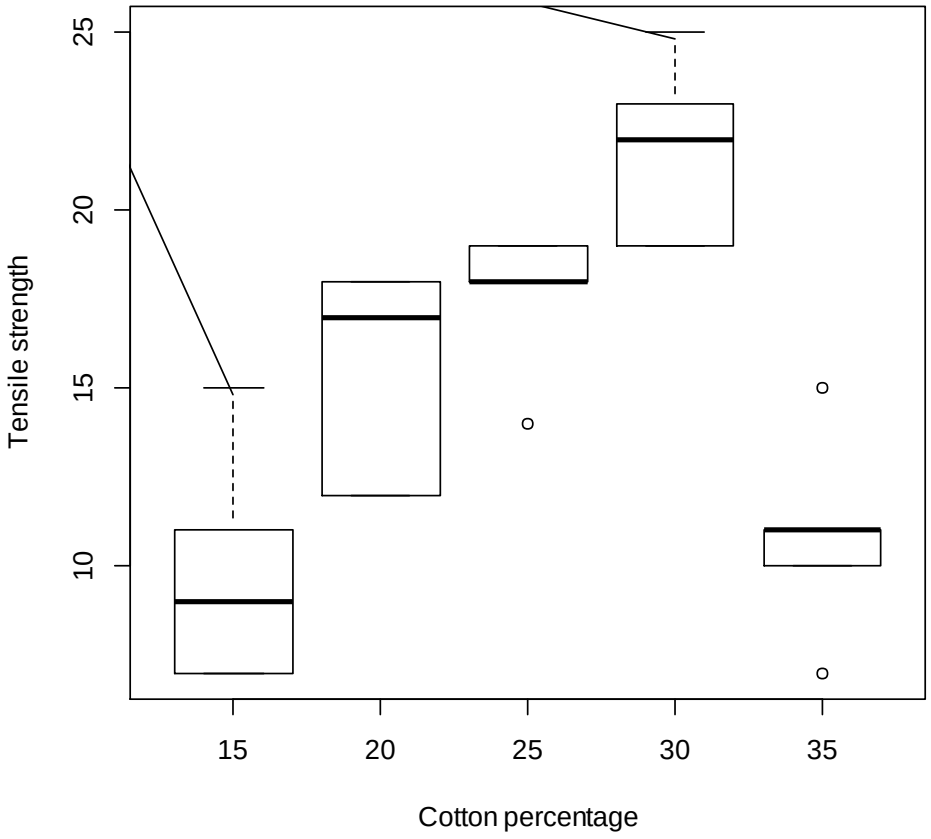
Test Sequences	Run Numbers	Percentage of Cotton
1	8	20
2	18	30
3	10	20
…		
…		
24	19	30
25	3	15

Data

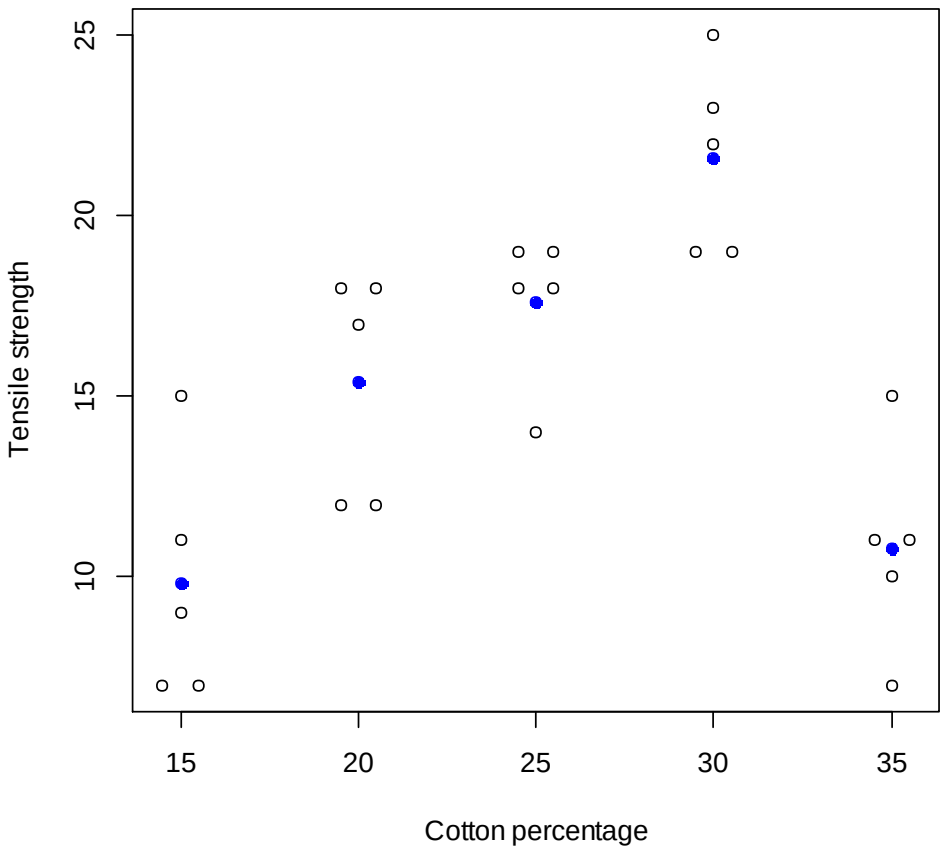
for the Cotton Percentage Example

Cotton Percentage	Observations					Total	Average
	1	2	3	4	5		
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						376	15.04

Tensile Strength Vs Cotton Percentage



Box plots



Scatter diagram

Exploratory Results

- We strongly suspect that:
  - Cotton content affects tensile strength
  - Around 30% cotton would result in maximum strength
- A more objective analysis:
  - Should we perform a *t*-test on all possible pairs of means?
  - No, results in a substantial increase in the type I error
- The appropriate procedure for testing the equality of several means is the analysis of variance (**ANOVA**)

ANOVA

- The name ANOVA is derived from a partitioning of total variability into its component parts: ANalysisOfVAriance
- The total sum of squares is a measure of overall variability in the data:
  - $SS_{total} = \sum_{i=1}^{a} \sum_{j=1}^{n} (yij - \overline{y}_{..})^2$
- The total variability in the data can be partitioned into a sum of squares of the differences *between* the treatment averages and the grand average, PLUS a sum of squares of the differences of observations *within* treatments from the treatment average:
  - $SS_{total} = SS_{treatment} + SS_{within}$

ANOVA

Degrees of freedom

- There are $N = an$ total observations, so SS_{total} has N -1 degrees of freedom.
- There are a levels of the factor, so $SS_{treatment}$ has a - 1 degrees of freedom.
- Thus, we have $N - a$ degrees of freedom for SS_{within}

ANOVA

- The analysis of variance identity provides us with two estimates of σ^2 – one based on the inherent variability within treatments and one based on the variability between treatments
  - $SS_{total} = SS_{treatment} + SS_{within}$
- If there are no differences in the treatment means, then these two estimates should be very similar.
- If they are not, we suspect that the observed differences must be caused by differences in treatment means.

ANOVA

for the Cotton Percentage Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
	$SS_{treatment}$			Test Statistics
Cotton Percentage	475.76	4	118.94	F_0=14.76
Error	SS_{within} 161.20	20	8.06	
Total	636.96	24		

ANOVA

Additional Concepts

- A factor is *fixed* if the levels of a factor are predetermined and the experimenter is interested only in those particular levels (e.g. 250, 300, or 350°F).
- A factor is classified as *random* if the levels are selected at random from a population of levels (e.g. 254, 287, and 326°F).
- When there is only one factor, the classification does not have any effect on how the data are analyzed.

Fixed Effects Model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

- y_{ij}: the ijth observation
- μ: grand mean, a parameter common to all treatments
- τ_i: treatment effect, a parameter unique to the ith treatment,
- ϵ_{ij}: random error

Least Squares Estimate

The Sum of squares of error :

$$L = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \mu - \tau_i)^2$$

Choose $\hat{\mu}$ and $\hat{\tau}$ that minimize L:

$$\left.\frac{\partial L}{\partial \mu}\right|_{\hat{\mu},\hat{\tau}_i} = 0$$

$$\left.\frac{\partial L}{\partial \tau}\right|_{\hat{\mu},\hat{\tau}_i} = 0$$

Least Squares Estimate

Adding a constraint:

$$\sum_{i=1}^{a} \hat{\tau}_i = 0$$

The least squares estimate for μ, τ_i :

$$\hat{\mu} = \bar{y}_{..}$$

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}, i = 1, 2, \ldots, a$$

Cotton Percentage example

-continued-

Estimate of the overall mean:

$$\hat{\mu} = \hat{y}_{..} = \frac{376}{25} = 15.04$$

Cotton Percentage example

-continued-

Estimate of treatment effects:

- $\widehat{\tau_1} = \overline{y_{1.}} - \overline{y_{..}} = 9.80 - 15.04 = -5.24$
- $\widehat{\tau_2} = \overline{y_{2.}} - \overline{y_{..}} = 15.40 - 15.04 = 0.36$
- $\widehat{\tau_3} = \overline{y_{3.}} - \overline{y_{..}} = 17.60 - 15.04 = -2.56$
- $\widehat{\tau_4} = \overline{y_{4.}} - \overline{y_{..}} = 21.60 - 15.04 = 6.56$
- $\widehat{\tau_5} = \overline{y_{5.}} - \overline{y_{..}} = 10.80 - 15.04 = -4.24$

- *see slide 7 for the data

Confidence Interval (CI)

$100(1-\alpha)$ percent CI on the ith treatment mean:

$$\overline{y_{i.}} \pm t_{\alpha/2,N-a}\sqrt{MS_E/n}$$

$100(1-\alpha)$ percent CI on the difference of two treatments mean:

$$\overline{y_{i.}} - \overline{y_{j.}} \pm t_{\alpha/2,N-a}\sqrt{2MS_E/n}$$

Cotton Percentage example

-continued-

95% CI on the mean of treatment 4:

$$\left[21.60 \pm (2.086)\sqrt{8.06/5}\right]$$

$\overline{y_{4.}}$ $t_{0.025,20}$ MS_E n

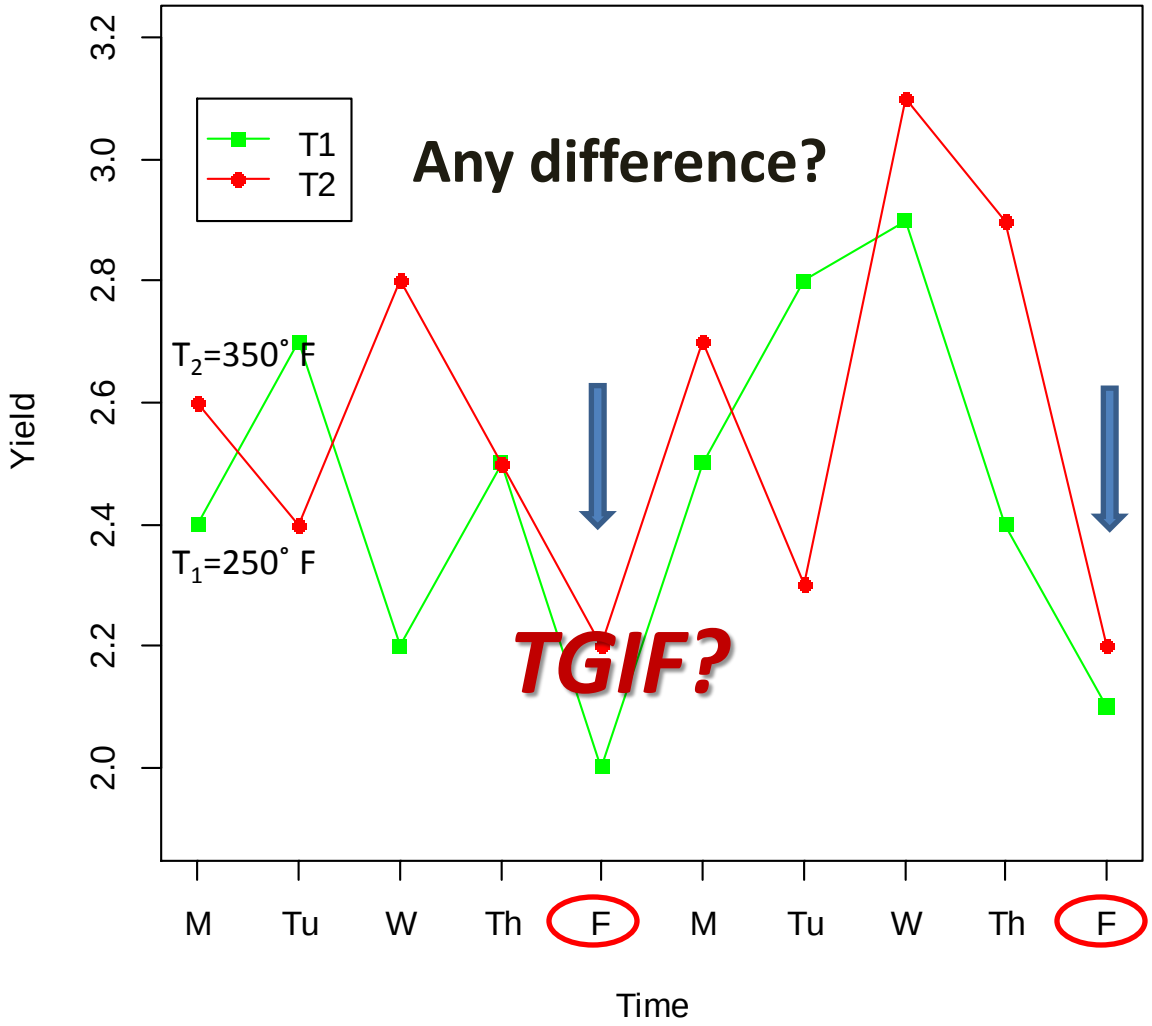
Temperature Example

- Determine the effects of temperature on process yields
  - Case I: Two levels of temperature setting
  - Case II: Three levels of temperature setting

Temperature Vs Process yields

		Temperature		
	Day	250 °F	300 °F	
Week # 1	M	2.4	2.6	Week #3
	Tu	2.7	2.4	
	W	2.2	2.8	
	Th	2.5	2.5	
	F	2.0	2.2	
Week # 2	M	2.5	2.7	Week # 4
	Tu	2.8	2.3	
	W	2.9	3.1	
	Th	2.4	2.9	
	F	2.1	2.2	

Exploratory analysis: Time Sequence plot



Analysis

t-test

- General form of t Statistics for one population

$$t = \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}}$$

θ: the parameter to be estimated

$\hat{\theta}$: the sample statistic that is the estimate of θ

$s_{\hat{\theta}}$: the estimator of the standard deviation of $\hat{\theta}$

Temperature Data I
t-test

Our interest : the difference $\mu_1 - \mu_2$

Let

$$\hat{\theta} = \overline{X_1} - \overline{X_2}$$

$$\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

- Generally, σ_1, σ_2 unknown, use estimates:
  - Two methods of estimating $\sigma_{\overline{X_1}-\overline{X_2}}$, yielding *exact* and *approximate* t-test.

Exact t-test

$$\hat{\theta} = \overline{X_1} - \overline{X_2}$$

Exact t-test:

zero or a constant

$$t = \frac{\overline{X_1} - \overline{X_2} - d}{s_p\sqrt{1/n_1 + 1/n_2}}$$

Estimate of $\sigma_{\overline{X_1}-\overline{X_2}}$ in the form of pooled variance

where n_1 and n_2 are sample sizes, $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$,

$s_a^2 = \frac{1}{n_a-1}\sum_{i=1}^{n_a}(x_{ai} - \overline{x_{a.}})^2$, $a = 1, 2$

Temperature Data I

exact t-test

For the temperature data:

$$T=\frac{(2.45-2.57)-0}{0.3005\sqrt{1/10+1/10}}=-0.893$$

$$P(t \leq -0.893)=0.191$$

p-value = 0.191, at 5% significance level, *fail to reject* H_0

Conclusion -- two scenarios:

- $\mu_1 \cong \mu_2$ OR
- Highly variable data in each sample

Exact t-test assumptions

- Normality of population
- Independence of samples
- Independence of observations
- Small (n<30) and equal or similar sample sizes($n_1 \cong n_2$)
- *Equal variance* $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$

Alternative approach:

Confidence Interval for difference

- One sided test, such as $\mu_1 < \mu_2$

$$(\overline{X_1} - \overline{X_2}) \pm t_{\alpha,d.f.} s_{\overline{x_1}-\overline{x_2}}$$

- Two Sided test, such as $\mu_1 \neq \mu_2$

$$(\overline{X_1} - \overline{X_2}) \pm t_{\frac{\alpha}{2},d.f.} s_{\overline{x_1}-\overline{x_2}}$$

Bounds can be calculated and the decision can be made based on interval information

ANOVA for Temperature Data I

Source of Variations	d.f.	SS	MS	F
Temperature	a-1= 1	0.072	0.072	0.797
Within	an-a= 18	1.626	0.090	
Total	an-1= 19	1.698		

$SS_{treatment}$

SS_{within}

Test Statistics
Under H_0, F~ F(a-1,n-1),
p=0.3838

$$SS_{treatment} = \sum_{i=1}^{a} n(y_{i.} - \bar{y}_{..})^2$$

$$SS_{within} = \sum_{i=1}^{a} \sum_{j=1}^{n} (yij - \overline{y_{i.}})^2$$

Numerator of s_p in the exact t test!

$$SS_{total} = \sum_{i=1}^{a} \sum_{j=1}^{n} (yij - \overline{y}_{..})^2$$

a: levels of treatment (temperature), a=2
n: replication of each treatment, n=10

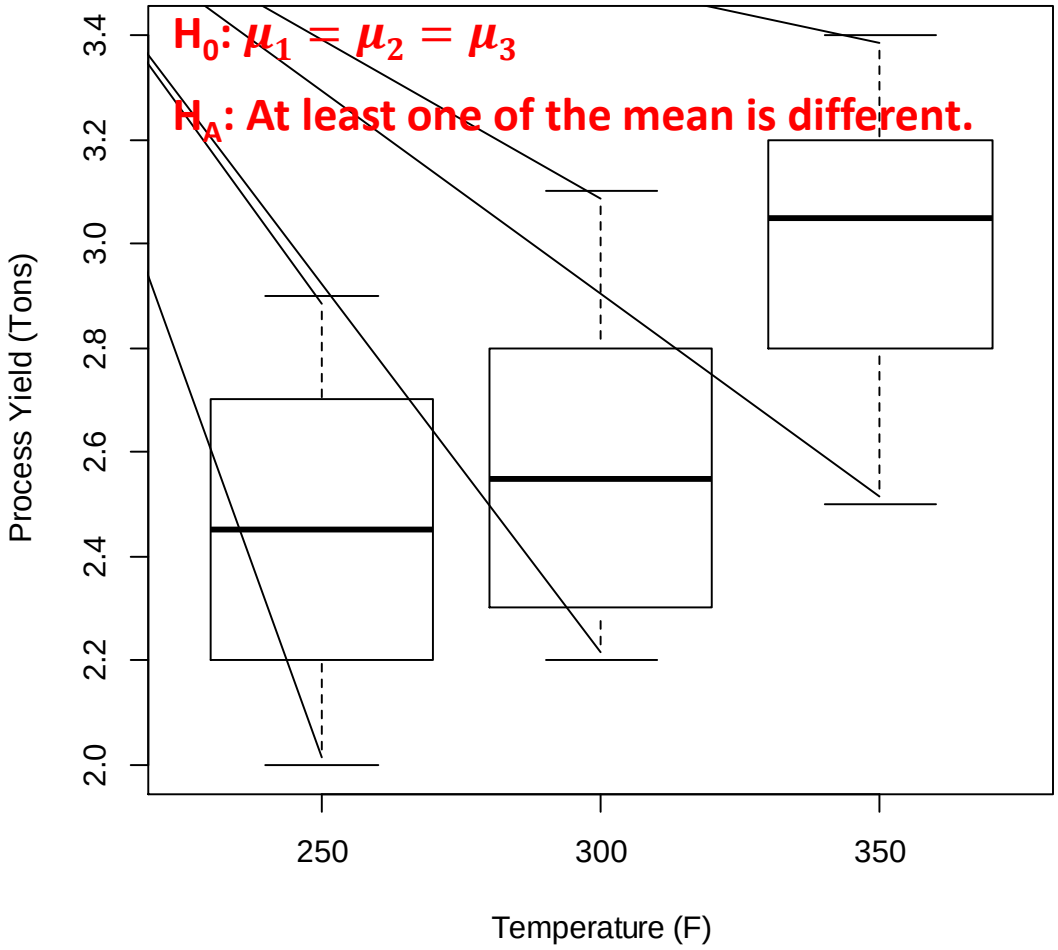
ANOVA uses SAME assumptions as t test!

Temperature Data II

Variations:
Week to Week?
Day to Day?

	Temperature		
Day	250 °F	300 °F	350°F
M	2.4	2.6	3.2
Tu	2.7	2.4	3.0
W	2.2	2.8	3.1
Th	2.5	2.5	2.8
F	2.0	2.2	2.5
M	2.5	2.7	2.9
Tu	2.8	2.3	3.1
W	2.9	3.1	3.4
Th	2.4	2.9	3.2
F	2.1	2.2	2.6

Exploratory Analysis: Temperature Data II- box plots



variability within each setting about the same

ANOVA for Temperature Data(3 levels)

Source of Variations	d.f.	SS	MS	F
Temperature	2	1.545	0.7725	8.91
Within	27	2.342	0.0867	**p-value=.001**
Total	29	3.887		**Reject H_0**

$$SS_{temp} = \sum_{i=1}^{3} \frac{\left(\sum_{j=1}^{10} y_{ij}\right)^2}{n} - \frac{\left(\sum_{i=1}^{3}\sum_{j=1}^{10} y_{ij}\right)^2}{an}$$

$$SS_{total} = \sum_{i=1}^{3}\sum_{j=1}^{10} {y_{ij}}^2 - \left(\sum_{i=1}^{3}\sum_{j=1}^{10} y_{ij}\right)^2/an$$

$$SS_{within} = SS_{total} - SS_{temp}$$

Typical Data for a Single Factor Experiments

Treatment (level)	Observations				Totals	Averages
1	y_{11}	y_{12}	…	Y_{1n}	$y_{1.}$	$\overline{y_{1.}}$
2	y_{21}	y_{22}	…	y_{2n}	$y_{2.}$	$\overline{y_{2.}}$
.	.	.	…	.	.	.
i	y_{i1}	y_{i2}	…	y_{in}	$y_{i.}$	.
.	.	.	…	.	.	.
a	y_{a1}	y_{a2}	…	y_{an}	$y_{a.}$	$\overline{y_{a.}}$
					$y_{..}$	$\overline{y_{..}}$

a

n

Clinical Trial Example

- A clinical trial is designed to estimate the efficacy of an experimental drug (D) compared to placebo (P) in congestive heart failure (CHF)
- Efficacy measure: the rate of change per week in distance walked after being administered therapy (D or P)
  - Baseline measure of left ventricular ejection fraction (LVEF) --- the lower the LVEF, the more serious CHF
  - 2 investigators
- Design and Analysis Tandem

Rate of Change in Distance Walked per Week by Drug, Investigator, and Baseline LVEF

Investigator	LVEF≤15		15<LVEF≤25		25<LVEF≤30		30<LVEF	
	D	P	D	P	D	P	D	P
1	0	69	842	-451	319	-178	165	-141
	56	1276	107	-132	-59	8	-342	-533
	327	-20	-131		1075	191	173	-244
	-255	-109			1173	-181		
		-107			-5	-220		
2	-144	-248	228	-84	168	-286	-59	-383
	604	-71	-657	-316	885	155	291	-631
	1221	144	-211	-90	745	30	0	-397
	497		75	37		-123	540	10
	168		-96					-27

Analysis of CHF data

- Means, standard deviations and sample sizes:
  - $\overline{X_D} = 240.6, S_D = 451.2, n_D = 32,$
  - $\overline{X_P} = -98.5, S_P = 322.4, n_p = 31,$

- The two-sample t with d.f. = 61 is

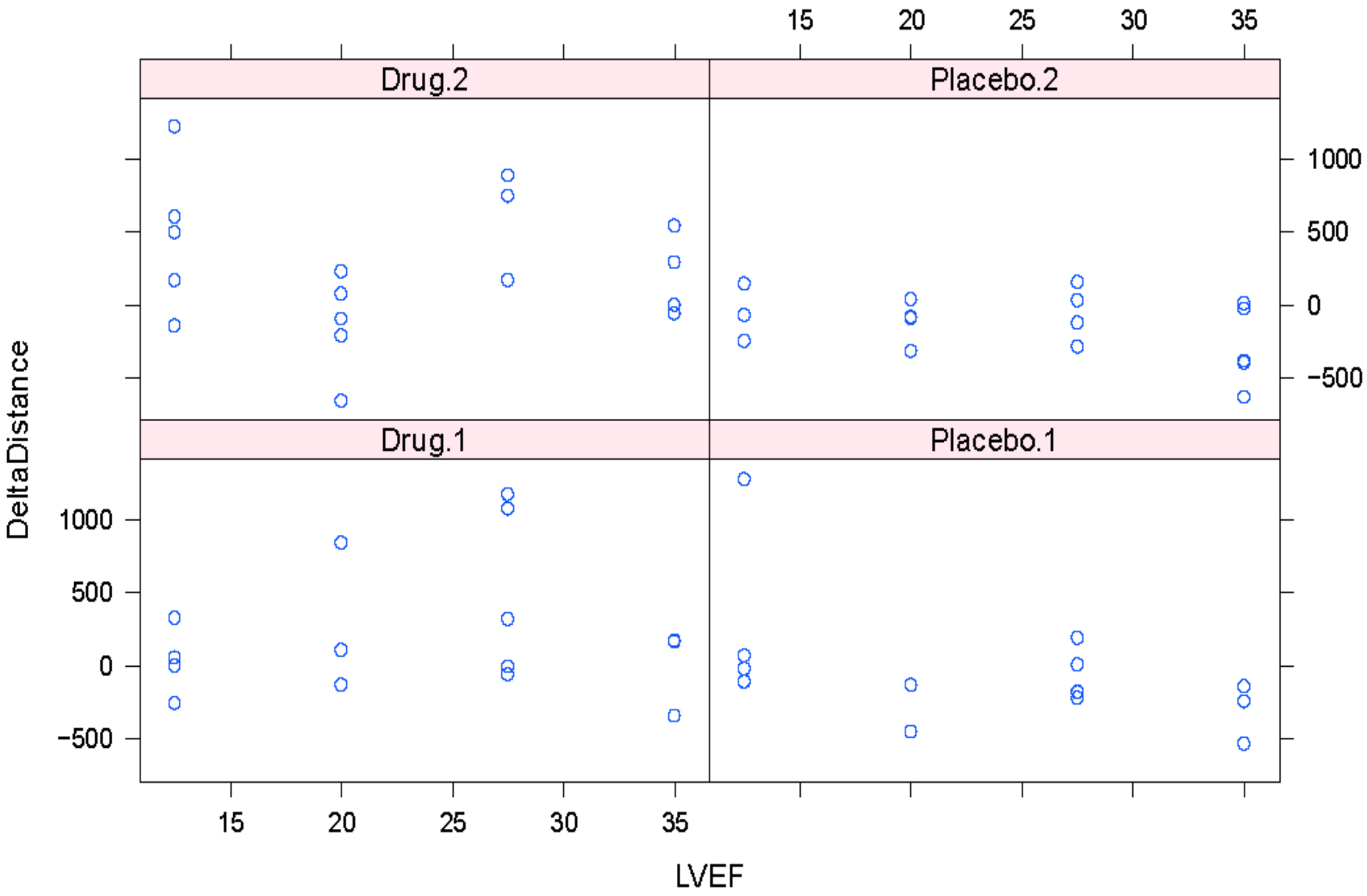
$$T = \frac{\overline{X_D} - \overline{X_P}}{\sqrt{{S_D}^2/n_D + {S_P}^2/n_P}} = 3.43$$

corresponds to a p-value = 0.001 : *reject $H_{0,}$ drug is significantly more efficacious than placebo* ***overall***

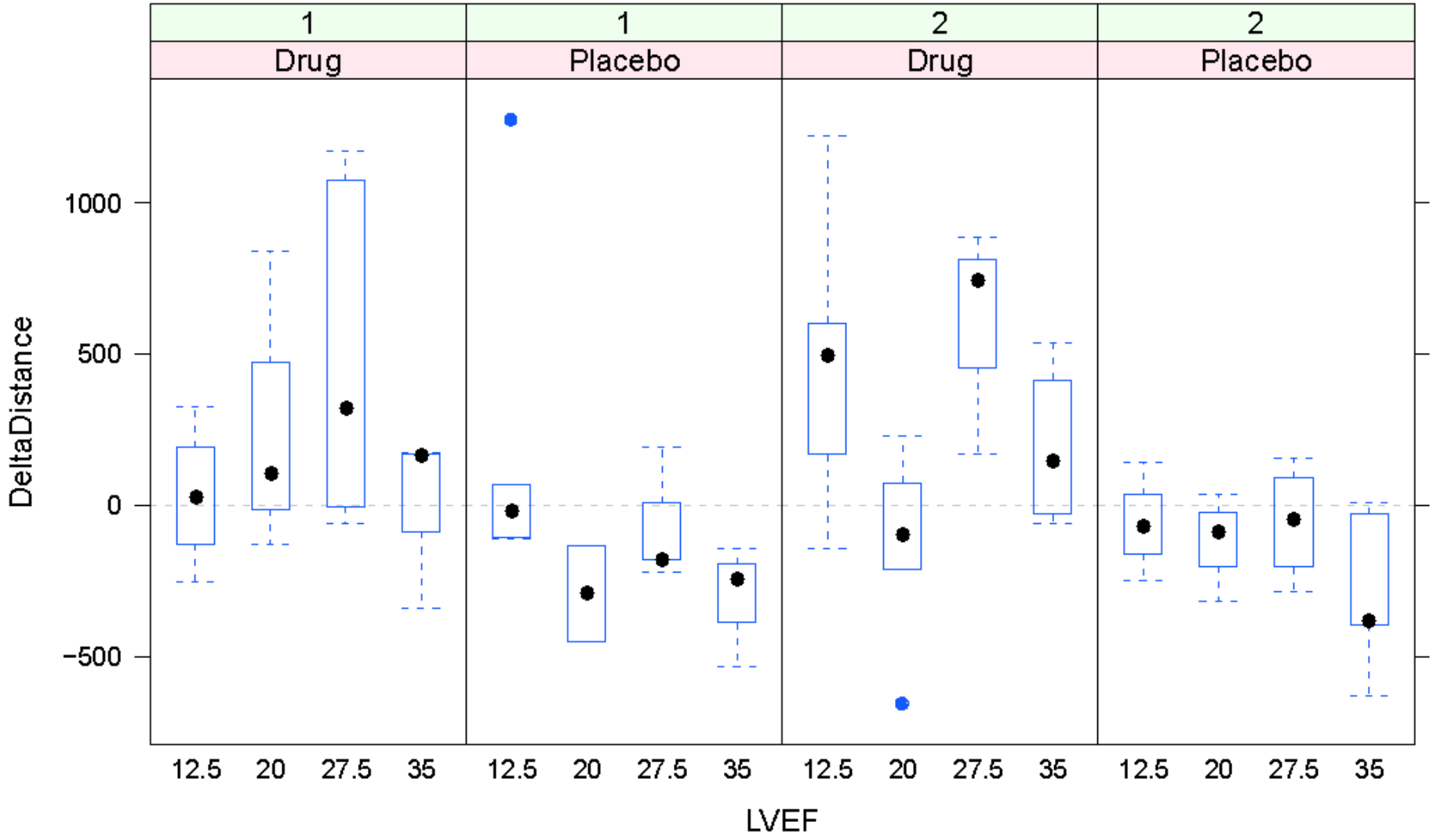
ANOVA for CHF data

		Original data		Rank transform	
Factor	d.f.	F	P	F	P
Drug	1	11.66	0.001	15.96	0.0002
Investigator	1	0.09	0.77	0.02	0.9
LVEF	1	1.48	0.23	1.18	0.28
Inv. * drug	1	0.21	0.65	0.06	0.81
LVEF*drug	1	1.76	0.19	2.63	0.11

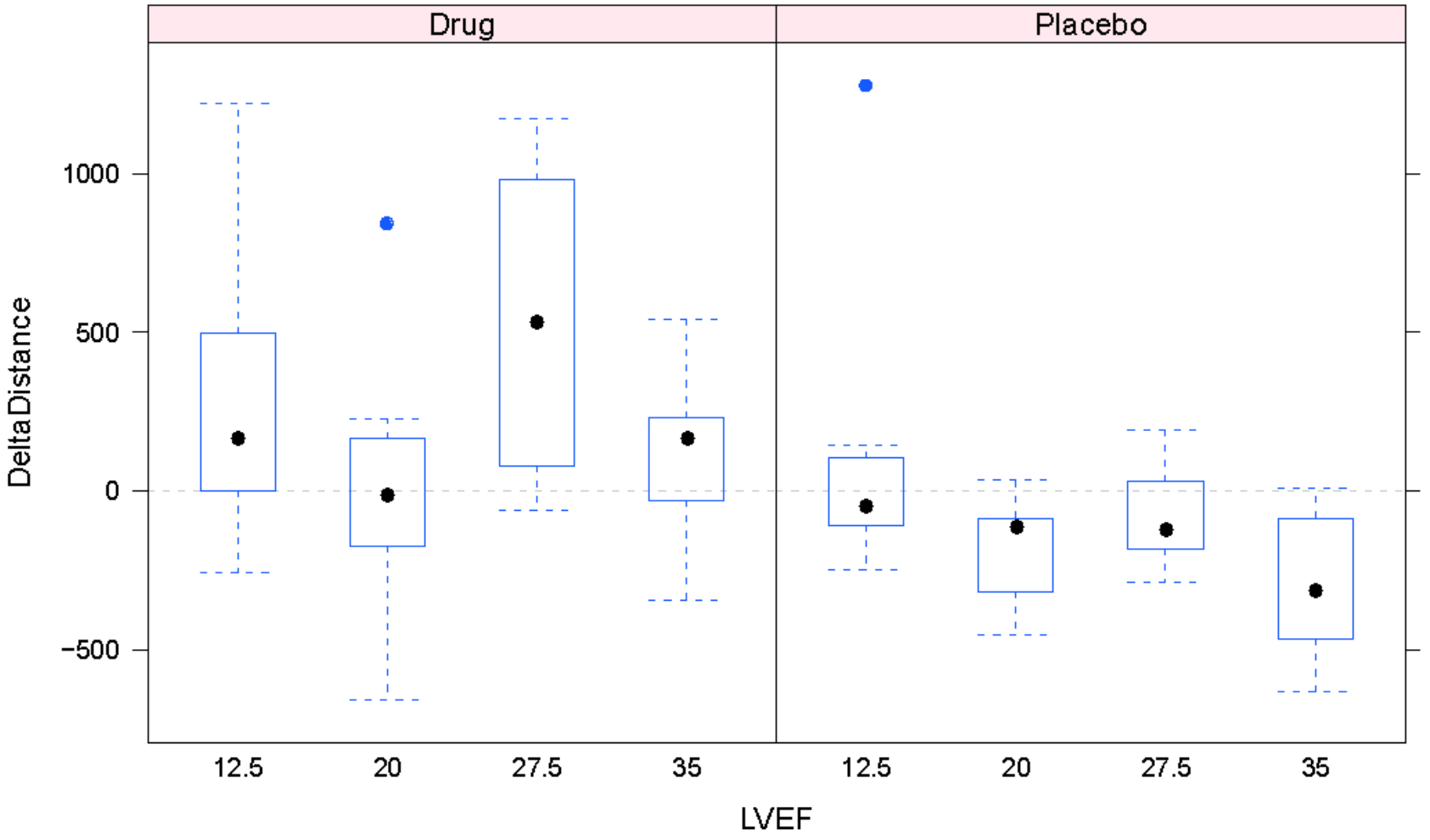
Exploratory Data Analysis



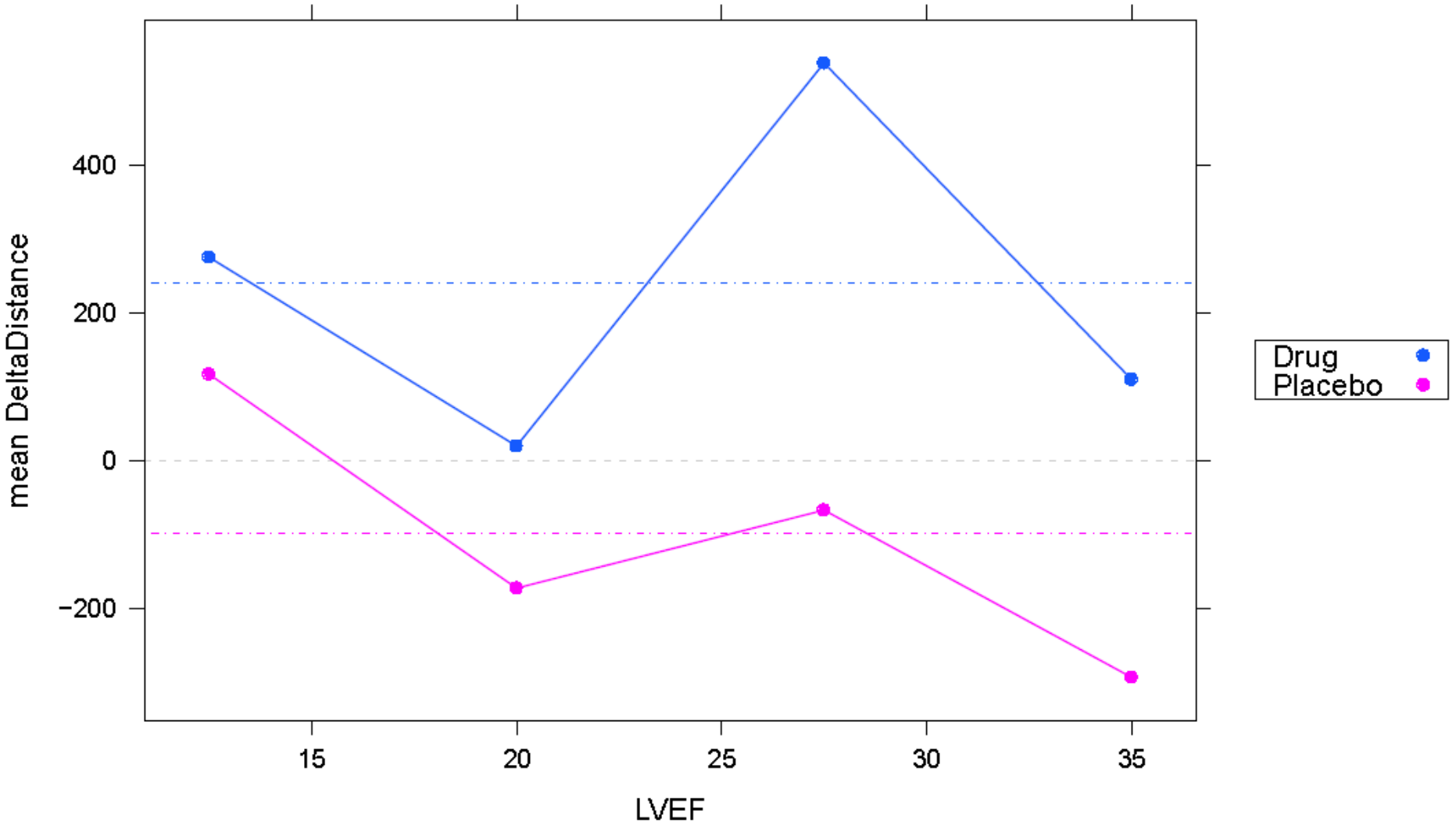
Box plots by Investigator



Box plots – Investigator effect pooled



Mean Distance by LVEF category



Mean Distance (D-P) by LVEF category