Comparisons Among Several Samples

SEVERAL-GROUPS PROBLEM

ANOVA

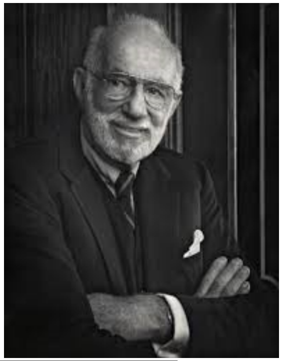
From Two-Groups to Many-Groups

- Subjects in a study can be in many different groups. This is known by two, equivalent terms:
  1. **SEVERAL-GROUP PROBLEM** (the two-sample tools from Chapters 2-4 are examples where there are two groups)
  2. **ONE-WAY CLASSIFICATION PROBLEM** (This naming convention extends to two-way classification, where there are two different grouping variables)
- When there are many different groups, there are many possible comparisons that can be made

Assumptions for the Several-Groups Problem

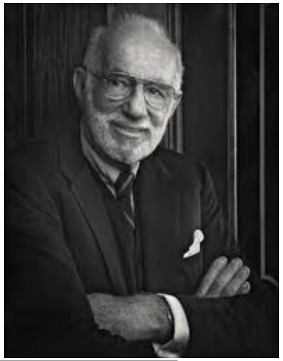
1. **Normality**: Each group are drawn from a normal distribution
2. **Equal population variances**: All the groups have the same standard deviation
3. **Independence**: The observations within and between each group are independent

Spock Trial



- 1968: Dr. Ben Spock was accused of conspiracy to violate the Selective Service Act by encouraging young men to resist being drafted into military service
- Jury Selection: A “venire” of 30 potential jurors are selected at random from a list of 300 names that were previously selected at random
- A jury is then selected not at random by the attorneys trying the case
- For this case, the venire consisted of only one woman, who was let go by the prosecution thus resulting in an all male jury.
- There was reason to believe that women were more sympathetic to Dr. Spock
- The defense argued that the judge in this case had a history of venires that underrepresented women, which is contrary to the law
- Let’s see if there is any evidence for this claim

Spock Trial



- To test the claim, the Spock Judge's (which we will call S) recent venires are compared with 6 other Judge's recent venires (which we notate A to F)

(Worth considering: How were these judges chosen?)

- There are two key questions
  1. Is there evidence that women are unrepresented on S's venire relative to A through F's?
  2. Is there evidence of a difference in women's representation on A to F's venires?

- The question of interest is addressed by 1

- The strength of the result in 1. would be diminished if 2 is true

Several-Groups Parameters

We will notate the population means of each group as $\mu_1,\ \mu_2, \dots,\ \mu_I$

(So, there are I different groups)

There is an additional standard deviation parameter σ

Hence, there are $I + 1$ parameters to estimate

<u>Spock Trial Example</u>: There are $\overbrace{\text{S, A, B, C, D, E, and F}}^{\text{(7-Groups)}}$ groups:

$\rightarrow$ There are $I + 1 = 7 + 1 = 8$ parameters

Estimation in Several-Groups Model

- Like usual, we will estimate population means with sample averages:

$\mu_1,\ \mu_2, \dots ,\ \mu_I \ \rightarrow\ \bar{Y}_1,\ \bar{Y}_2, \dots , \bar{Y}_I$

We can also form a pooled estimate of σ using a weighted average of each group's sample standard deviation, $s_1,\ s_2, \dots ,\ s_I$

$$\boxed{s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_I - 1)}.}$$

Note: The denominator equals *n* - I and hence degrees of freedom = *n* - I

Reminder:
Std. Dev. of the Difference

Suppose we want to test the difference in means between group i and group j: $\mu_i - \mu_j$

We estimate this difference: $\mu_i - \mu_j \rightarrow \bar{Y}_i - \bar{Y}_j$

We can compute the standard deviation of $\bar{Y}_i - \bar{Y}_j$:

(EQUAL VARIANCES)

$$SD(\bar{Y}_i - \bar{Y}_j) = \sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}} = \sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \rightarrow SE(\bar{Y}_i - \bar{Y}_j) = s_p\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Note: We used information from all groups to test this difference!

This s_p is **not** exactly the same as for pooled two-sample t-test

Spock Data Steps

```
DATA spock;
      INPUT percFemale judge $;
      DATALINES;
06.4 S
08.7 S
13.3 S
13.6 S
15.0 S
15.2 S
17.7 S
18.6 S
23.1 S
16.8 A
```

Question: Suppose we wish to test if the "S" judge's venires are different from the "F" judge's.

```
DATA spockVsF;
    SET spock;
    if (judge NE 'S') & (judge NE 'F')  THEN DELETE;
RUN;
```

Two Judge Analysis w/ t-Tools

```
PROC TTEST DATA = spockVsF ORDER=DATA;
    CLASS judge;
    VAR percFemale;
RUN;
```

judge	N	Mean	Std Dev	Std Err	Minimum	Maximum
S	9	14.6222	5.0388	1.6796	6.4000	23.1000
F	9	26.8000	5.9689	1.9896	16.5000	36.2000
Diff (1-2)		-12.1778	5.5234	2.6038		

judge	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
S		14.6222	10.7491	18.4954	5.0388	3.4035	9.6532
F		26.8000	22.2119	31.3881	5.9689	4.0317	11.4350
Diff (1-2)	Pooled	-12.1778	-17.6975	-6.6580	5.5234	4.1137	8.4063
Diff (1-2)	Satterthwaite	-12.1778	-17.7102	-6.6454			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	16	-4.68	0.0003
Satterthwaite	Unequal	15.562	-4.68	0.0003

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	8	1.40	0.6431

Statistical Conclusion: We find that there is substantial evidence that the difference in the mean percentage of females on judge S and judge F venires is not equal to zero.

Estimated Diff = -12.1778
Pooled Std. Error = 2.6038
t-Statistic = -4.68
Deg. of freedom = 16

Two Judge Analysis w/ Several-Groups

From PROC TTEST:
Estimated Diff = -12.1778
Pooled Std. Error = 2.6038
t-Statistic = -4.68
Deg. of freedom = 16

Deg. of freedom = 46 – 7 = 39

The GLM Procedure
Dependent Variable: percFemale

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1927.080865	321.180144	6.72	<.0001
Error	39	1864.445222	47.806288		
Corrected Total	45	3791.526087			

R-Square	Coeff Var	Root MSE	percFemale Mean
0.508260	26.01027	6.914209	26.58261

Source	DF	Type I SS	Mean Square	F Value	Pr > F
judge	6	1927.080865	321.180144	6.72	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
judge	6	1927.080865	321.180144	6.72	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Estimate Spock judge to F judge	-12.1777778	3.25938944	-3.74	0.0006

```
PROC GLM DATA = spock ORDER=DATA;
      CLASS judge;
      MODEL percFemale = judge;
      ESTIMATE 'Estimate Spock judge to F judge' judge 1 0 0 0 0 0 -1;
RUN;
```

Two Judge Analysis: Conclusion

Question: Suppose we wish to test if the "S" judge's venires are different from the "F" judge's.

Answer: There is evidence that the means of the two groups are different

We can use regular t-Tools or several-group analysis.

The several-group analysis allows us to use all of the available information → larger degrees of freedom → smaller p-values (in general)

Note: In this particular case, it happened to have a smaller estimate of the standard deviation in the t-Tools case than the several-group one. This shouldn't be expected to happen in general

Several-Groups Analysis: Analysis of Variance (ANOVA)

We can do a lot more than reduce the degrees of freedom in a two-sample comparison

A core stage in the analysis of several-group data is answering the question "is there evidence that any of the groups are different?"

This question is answered by conducting an **ANALYSIS OF VARIANCE (ANOVA)**

(**Warning:** Though the word "variance" appears in the name, it is very much a test of means, not variances)

Analysis of Variance (ANOVA)

- Did the data come from groups that **all** had the same mean and standard deviation?

$\mu_1 = \mu_2 = \cdots = \mu_I = \mu$ (and σ) (The **REDUCED OR EQUAL MEANS MODEL**)

$\rightarrow \bar{Y}$ (The **GRAND MEAN**)

- Did the data come from groups that **do not all** have the same mean but still have the same standard deviation?

$\mu_k \neq \mu_l$ for some k, l (and σ) (The **FULL OR SEPARATE MEANS MODEL**)

$\rightarrow \bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_I$

- The "Extra-Sum-of-Squares" principle allows us to compare the two competing models via **RESIDUALS**

Large residuals indicate that the model fits poorly.

		Equal means		Separate means	
Judge	**%W**	Est.	Res.	Est.	Res.
Spock	6.4	26.6	−20.2	14.6	−8.2
Spock	8.7	26.6	−17.9	14.6	−5.9
Spock	13.3	26.6	−13.3	14.6	−1.3
Spock	13.6	26.6	−13.0	14.6	−1.0
Spock	15.0	26.6	−11.6	14.6	0.4
Spock	15.2	26.6	−11.4	14.6	0.6
Spock	17.7	26.6	−8.9	14.6	3.1
Spock	18.6	26.6	−8.0	14.6	4.0
Spock	23.1	26.6	−3.5	14.6	8.5
A	16.8	26.6	−9.8	34.1	−17.3
A	30.8	26.6	4.2	34.1	−3.3
A	33.6	26.6	7.0	34.1	−0.5
A	40.5	26.6	13.9	34.1	6.4
A	48.9	26.6	22.3	34.1	14.8
B	27.0	26.6	0.4	33.6	−6.6
B	28.9	26.6	2.3	33.6	−4.7
B	32.0	26.6	5.4	33.6	−1.6
B	32.7	26.6	6.1	33.6	−0.9
B	35.5	26.6	8.9	33.6	1.9
B	45.6	26.6	19.0	33.6	12.0

In the equal-means model, estimated means are equal to the grand average.

In the separate-means model, estimated means are the group averages.

		Equal means		Separate means	
Judge	**%W**	Est.	Res.	Est.	Res.
C	21.0	26.6	−5.6	29.1	−8.1
C	23.4	26.6	−3.2	29.1	−5.7
C	27.5	26.6	0.9	29.1	−1.6
C	27.5	26.6	0.9	29.1	−1.6
C	30.5	26.6	3.9	29.1	1.4
C	31.9	26.6	5.3	29.1	2.8
C	32.5	26.6	5.9	29.1	3.4
C	33.8	26.6	7.2	29.1	4.7
C	33.8	26.6	7.2	29.1	4.7
D	24.3	26.6	−2.3	27.0	−2.7
D	29.7	26.6	3.1	27.0	2.7
E	17.7	26.6	−8.9	27.0	−9.3
E	19.7	26.6	−6.9	27.0	−7.3
E	21.5	26.6	−5.1	27.0	−5.5
E	27.9	26.6	1.3	27.0	0.9
E	34.8	26.6	8.2	27.0	7.8
E	40.2	26.6	13.6	27.0	13.2
F	16.5	26.6	−10.1	26.8	−10.3
F	20.7	26.6	−5.9	26.8	−6.1
F	23.5	26.6	−3.1	26.8	−3.3
F	26.4	26.6	−0.2	26.8	−0.4
F	26.7	26.6	0.1	26.8	−0.1
F	29.5	26.6	2.9	26.8	2.8
F	29.8	26.6	3.2	26.8	3.0
F	31.9	26.6	5.3	26.8	5.1
F	36.2	26.6	9.6	26.8	9.4

If the equal means model is correct, then the magnitude of the residuals should be about the same as the separate mean model

Important: as the residuals of the equal means model will always be larger in magnitude than separate means

Extra-Sum-of-Squares

To quantify “about the same as”, we can add up the residuals

However, positive and negative residuals both give equal amount of evidence, and if added up cancel each other out

Instead of adding the raw residuals, we **ADD UP THE SQUARED RESIDUALS**

ADDING UP THE SQUARED RESIDUALS $\leftrightarrow$ **RESIDUAL SUM OF SQUARES**

Procedure: Find the residual sum of squares of the equal and separate means models and compare them

If they are very different, this is evidence that separate means is a better model

Like usual, we appeal to probability theory to define “very different”

Extra-Sum-of-Squares

EXTRA SUM OF SQUARES = ESS =

residual sum of squares(reduced) - residual sum of squares(full)

Large extra sum of squares indicates the full model fits much better

To make the size of the extra sum of squares meaningful, we need to standardize it

F-statistic= $\dfrac{ESS/_{\text{df of } ESS}}{\hat{\sigma}_p^2}$

Here:

df of ESS = # parameters in full model $(I+1)$ - # parameters in reduced model (2)

$\hat{\sigma}_p^2 = \hat{\sigma}_{full}^2$ is the estimate of the variance based on the full model

F-statistic

If all of the means are equal, then:

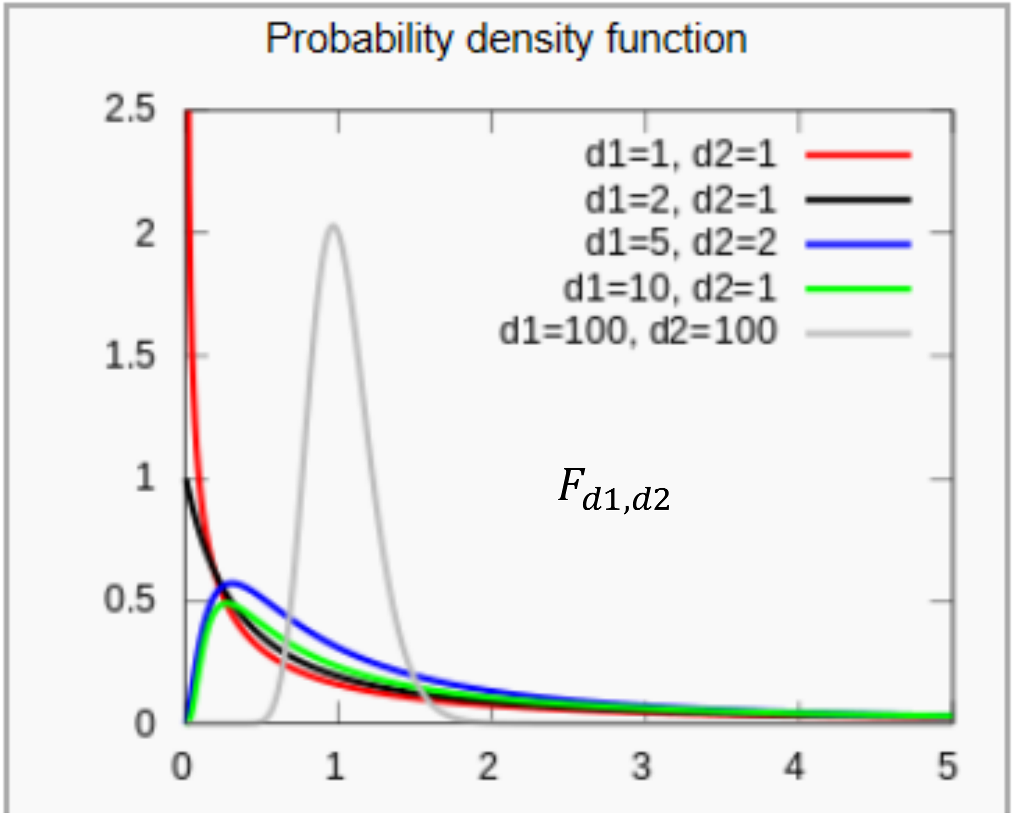
F-statistic $= \frac{ESS/_{\text{df of } ESS}}{\hat{\sigma}_p^2}$ follows an F-distribution

The F-distribution has a numerator degrees of freedom and a denominator degrees of freedom and is written

$F_{numerator,denominator}$

(Compare this with a t-distribution: t_{df})

F-Distributions



From Extra Sums of Squares to the ANOVA Table

Reminder

(ESS)

EXTRA SUM OF SQUARES = ESS =

residual sum of squares(reduced) - residual sum of squares(full)

RSS(reduced) RSS(full)

Large extra sum of squares indicates the full model fits much better

- Did the data come from groups that **all** had the same mean?

H_0: $\mu_1 = \mu_2 = \cdots = \mu_I = \mu$ (The **REDUCED OR EQUAL MEANS MODEL**)

- Did the data come from groups that **do not all** have the same mean

H_A: $\mu_k \neq \mu_l$ for some k, l (The **FULL OR SEPARATE MEANS MODEL**)

From Extra Sums of Squares to the ANOVA Table

Reminder

(ESS)
EXTRA SUM OF SQUARES =

residual sum of squares(reduced) - residual sum of squares(full)

RSS(reduced) RSS(full)

Large extra sum of squares indicates the full model fits much better

$F_{I-1,n-I}$

ANOVA table:

Source	DF	SS	MS	F	Pr > F
Model (Between)	$I-1$	ESS	ESS/$(I-1)$	F-statistic	p-value
Error (Within)	$n-I$	RSS(full)	RSS(full)/$(n-I)$		
Corrected Total (Total)	$n-1$	RSS(reduced)			

From Extra Sums of Squares to the ANOVA Table

Reminder

F-statistic= $\dfrac{ESS/_{\text{df of } ESS}}{\hat{\sigma}_p^2} = \dfrac{ESS/_{I-1}}{RSS(full)/_{n-I}}$

Here:

$\text{df of } ESS$ = # parameters in full model - # parameters in reduced model

$\hat{\sigma}_p^2$ is the estimate of the variance based on the full model

$F_{I-1,n-I}$

ANOVA table:

Source	DF	SS	MS	F	Pr > F
Model (Between)	$I-1$	ESS	ESS/$(I-1)$	F-statistic	p-value
Error (Within)	$n-I$	RSS(full)	RSS(full)/$(n-I)$		
Corrected Total (Total)	$n-1$	RSS(reduced)			

Class Example

Treatment 1	Treatment 2	Placebo
3	10	20
5	12	22
7	14	24

Class Example: SAS

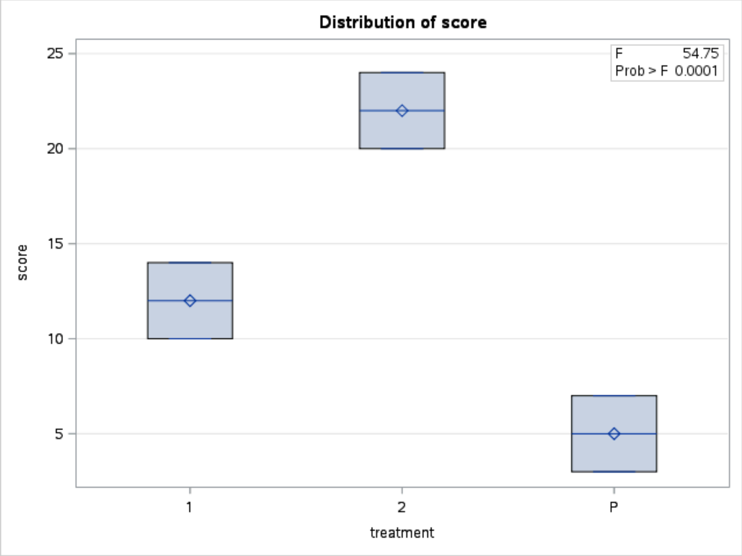
Ho: $\mu_1 = \mu_2 = \mu_3$ (Equal Means Model)
Ha: At least 1 pair are different (Separate Means Model)

```
DATA example;
    INPUT score treatment $;
    DATALINES;
3  P
5  P
7  P
10 1
12 1
14 1
20 2
22 2
24 2
;
RUN;
```

```
PROC GLM DATA = example;
    CLASS treatment;
    MODEL score = treatment;
RUN;
```

The GLM Procedure
Dependent Variable: score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	438.0000000	219.0000000	54.75	0.0001
Error	6	24.0000000	4.0000000		
Corrected Total	8	462.0000000			



ANOVA in the Spock Example

Ho: All means are equal (Spock and Others)

Ha: At least 2 are different (Spock and Others)

```
PROC GLM DATA = spock ORDER=DATA;
      CLASS judge;
      MODEL percFemale = judge;
RUN;
```

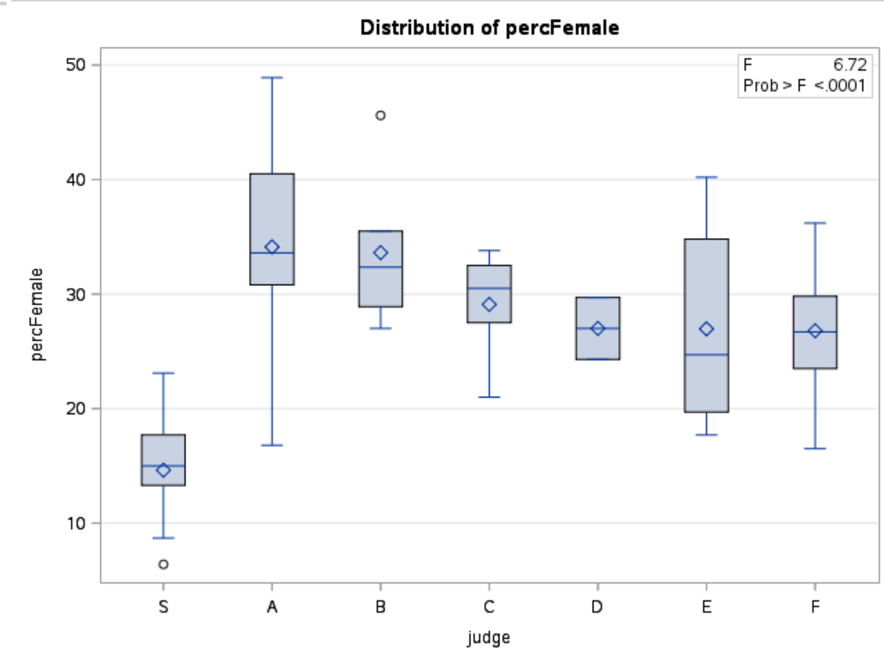
RSS(reduced)

ESS

RSS(full)

The GLM Procedure
Dependent Variable: percFemale

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1927.080865	321.180144	6.72	<.0001
Error	39	1864.445222	47.806288		
Corrected Total	45	3791.526087			



Statistical Conclusion: There is substantial evidence that the mean percent of females on venires is not the same for all seven judges (ANOVA p-value less than 0.0001).

Assumptions for ANOVA

- **Normality**: Each group are drawn from a normal distribution

  (Look at QQ plot or check sample sizes)

- **Equal Standard Deviations**: All the groups have the same standard deviation

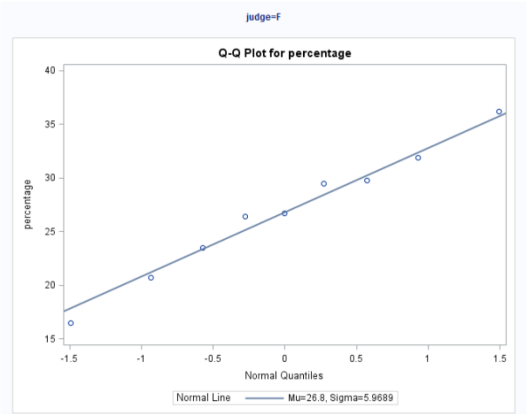
  (Look at the residuals)

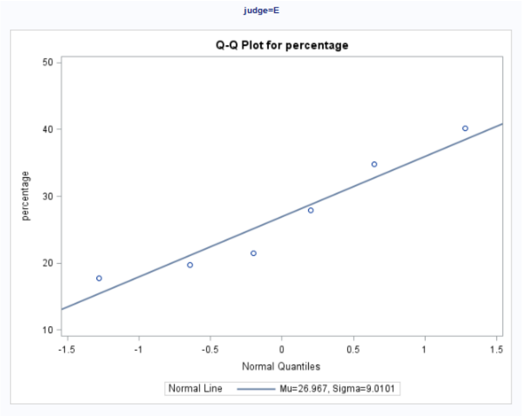
- **Independence**: The observations within and between each group are independent

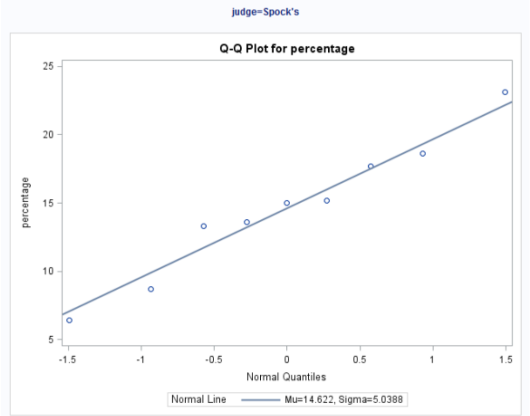
  (Examine the study to see if independence is reasonable)

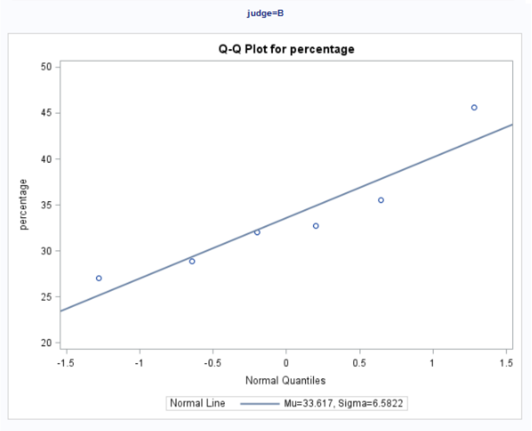
Normality

```
PROC SORT DATA = spock;
    BY judge;
RUN;
PROC UNIVARIATE DATA = spock;
    BY judge;
    VAR percFemale;
    QQPLOT / NORMAL(MU=EST SIGMA=EST L=2);
RUN;
```





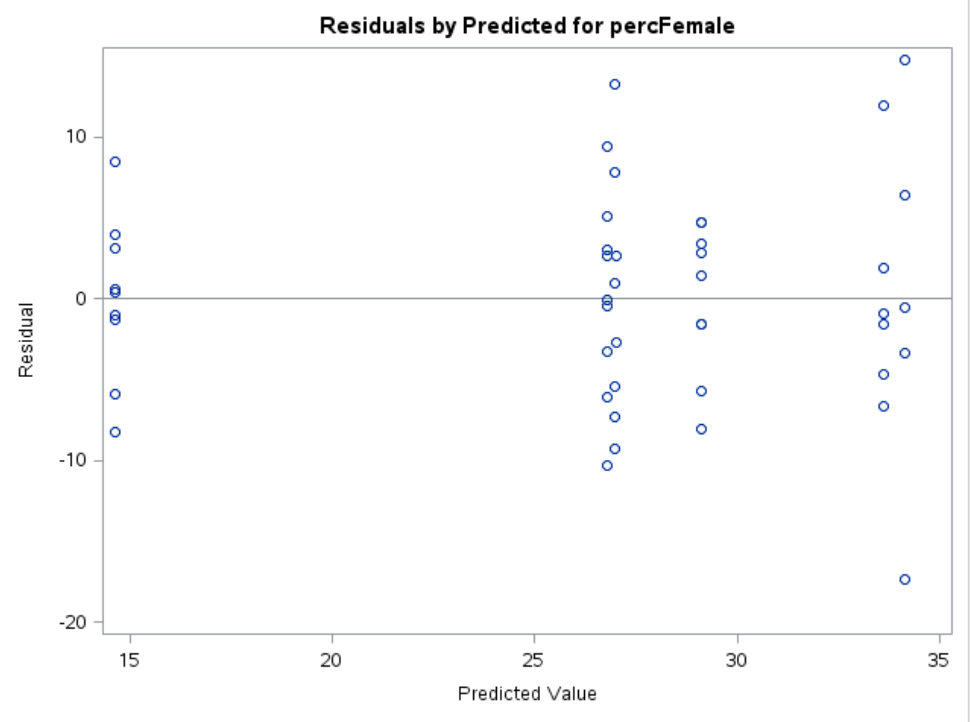




(This is only 4 out of 7 plots for brevity's sake)

Residuals

```
PROC GLM DATA = spock ORDER=DATA PLOTS(UNPACK)=DIAGNOSTICS;
      CLASS judge;
      MODEL percFemale = judge;
RUN;
```



We are looking for:

1. Funnel Shapes
2. Non-constant Variance

(no problems in this case)

Notes:

1. Only look at the first reported plot with this code.
2. "Predicted Value" is the same as "Estimated Means" in Display 5.15