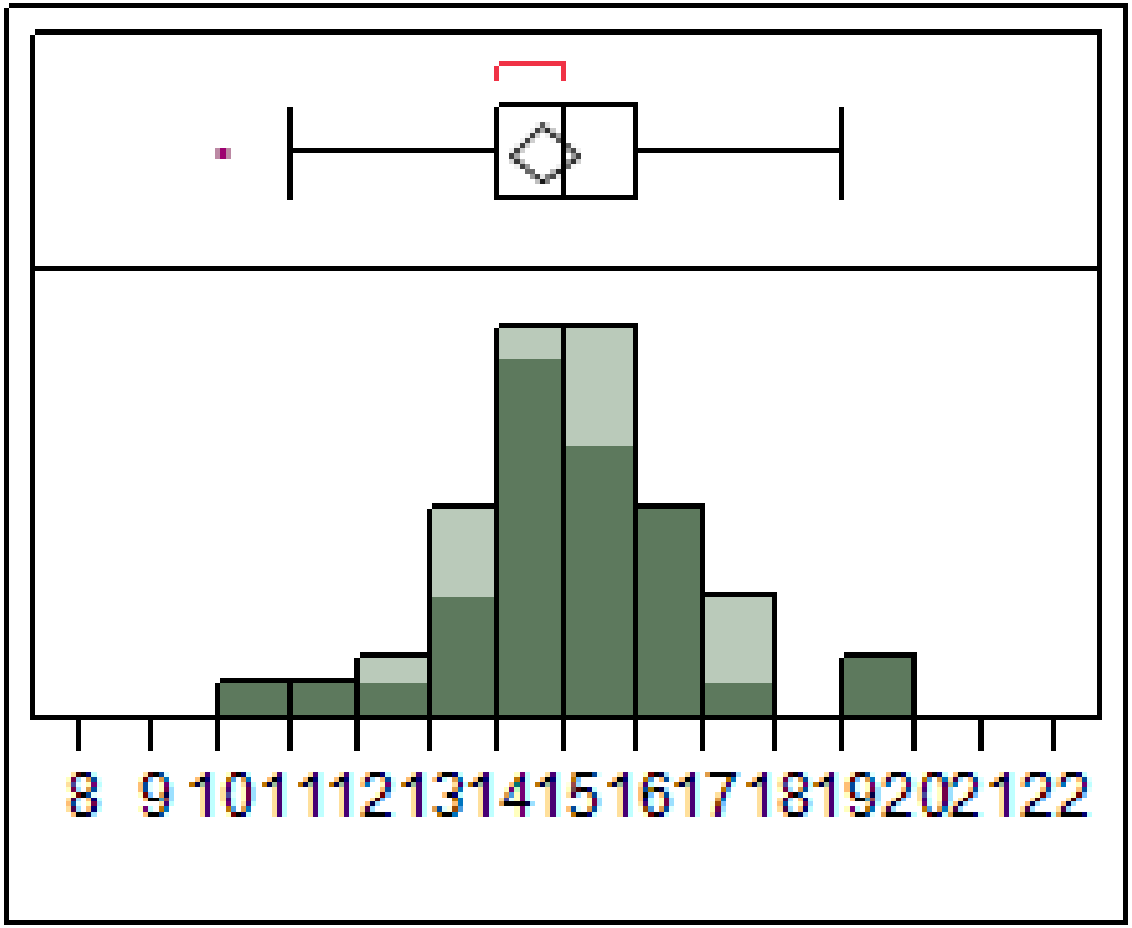
分散分析の基礎

国立国際医療研究センター
医学統計研究室
田中　紀子

営利目的でのご利用はご遠慮ください

分散

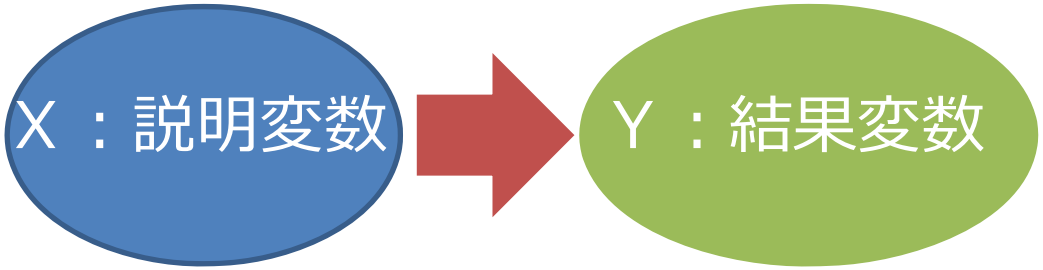


$$\sum (y-\bar{y})^2 / (n-1)$$

営利目的でのご利用はご遠慮ください

ICRweb http://www.icrweb.jp/

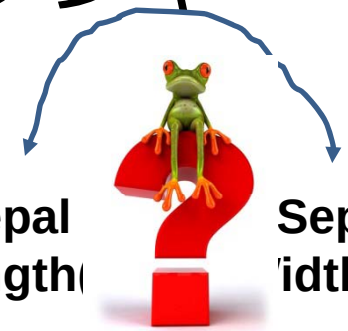
データを分析する



営利目的でのご利用はご遠慮ください

ICRweb http://www.icrweb.jp/

例：フィッシャーのあやめのデータ



id	Species (Y_1)	Sepal length(	Sepal width(x_2)	Petal length(x_3)	Petal Width(x_4)
1	Setosa	50	33	14	2
2	Virginica	64	28	56	22
…					
150	Versicolor	67	30	50	17

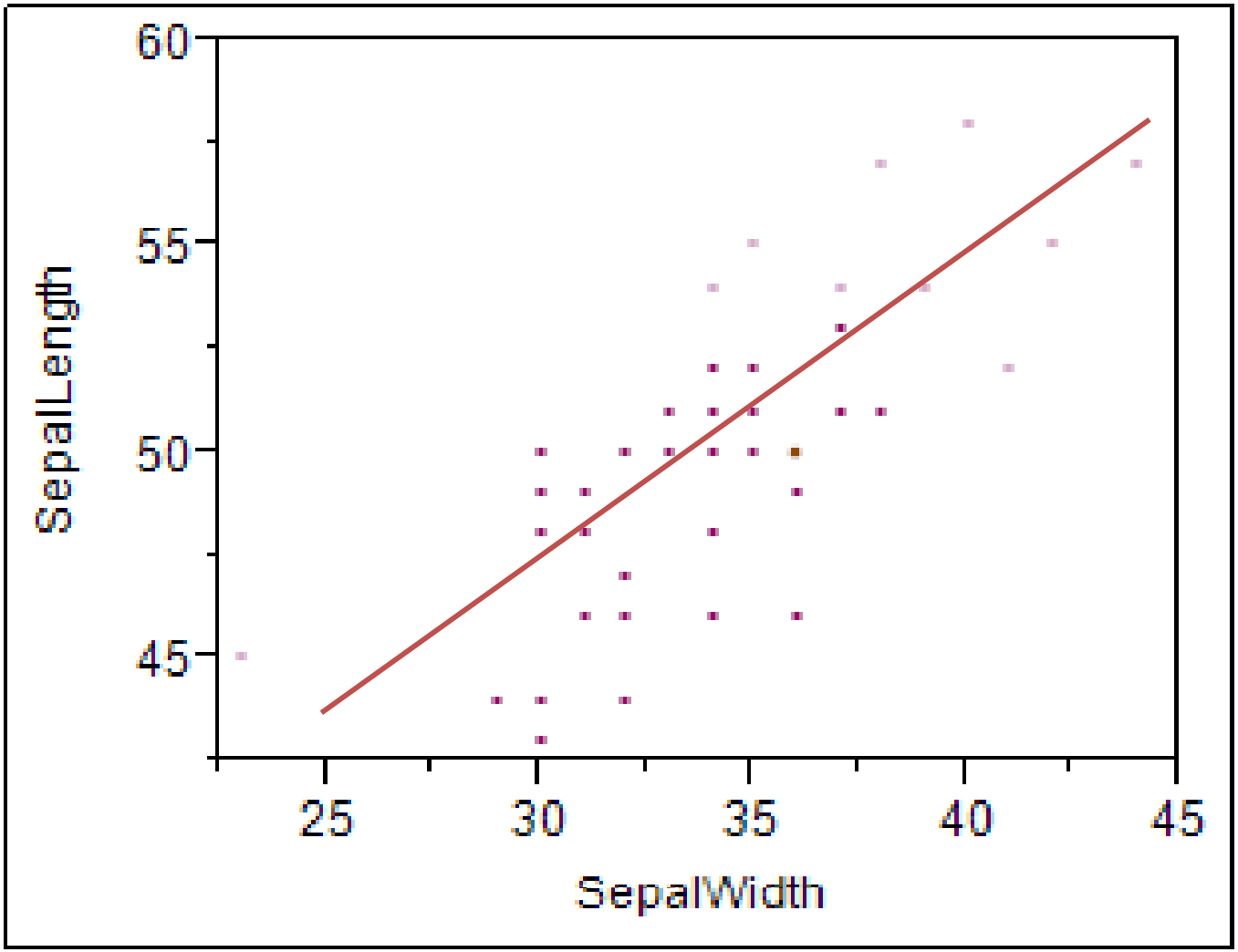
Setosa(n=50)

Virginica(n=50)

Versicolor(n=50)

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

がくの長さと幅の関係

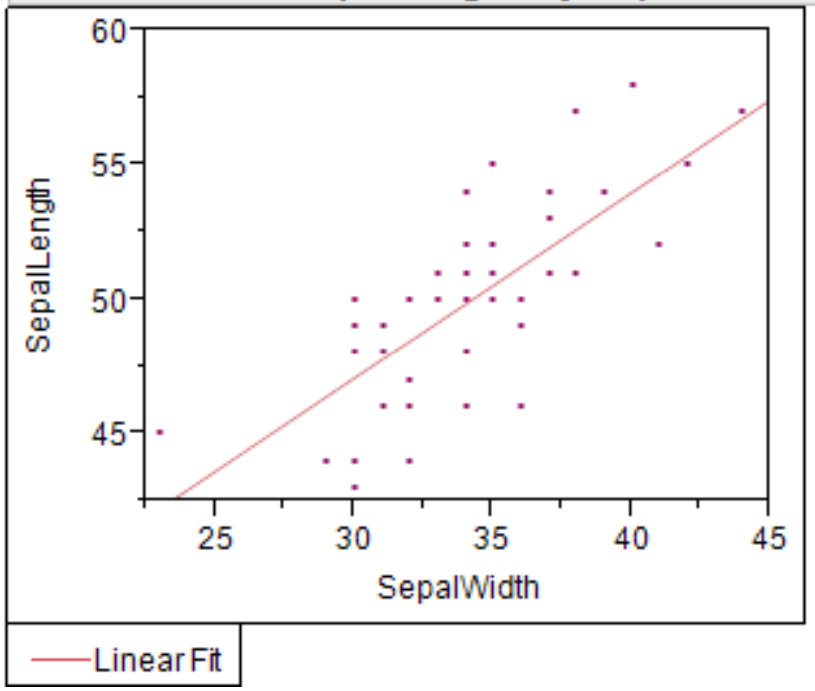


営利目的でのご利用はご遠慮ください

ICRweb http://www.icrweb.jp/

がくの幅と長さの関係

Bivariate Fit of SepalLength By SepalWidth Species=Setosa



Linear Fit

SepalLength = 26.390012 + 0.6904897*SepalWidth

Summary of Fit

RSquare	0.551376
RSquare Adj	0.542029
Root Mean Square Error	2.385422
Mean of Response	50.06
Observations (or Sum Wgts)	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	335.68848	335.688	58.9937
Error	48	273.13152	5.690	Prob > F
C. Total	49	608.82000		<.0001*

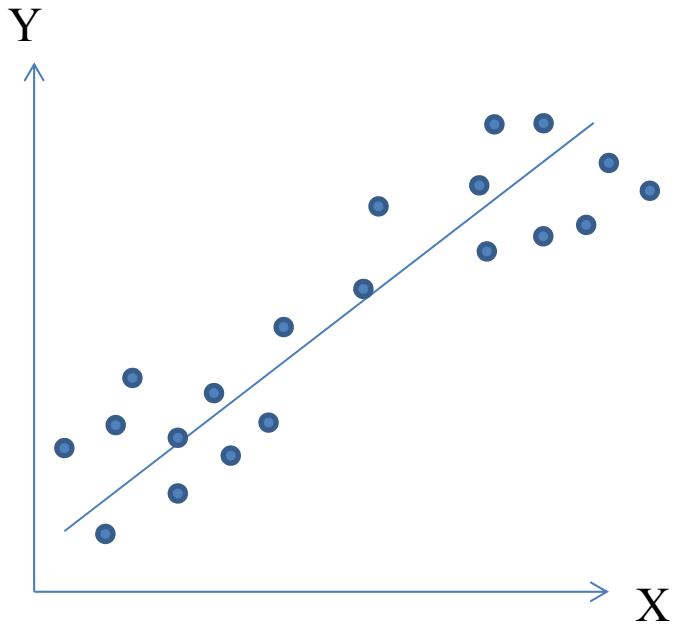
Parameter Estimates

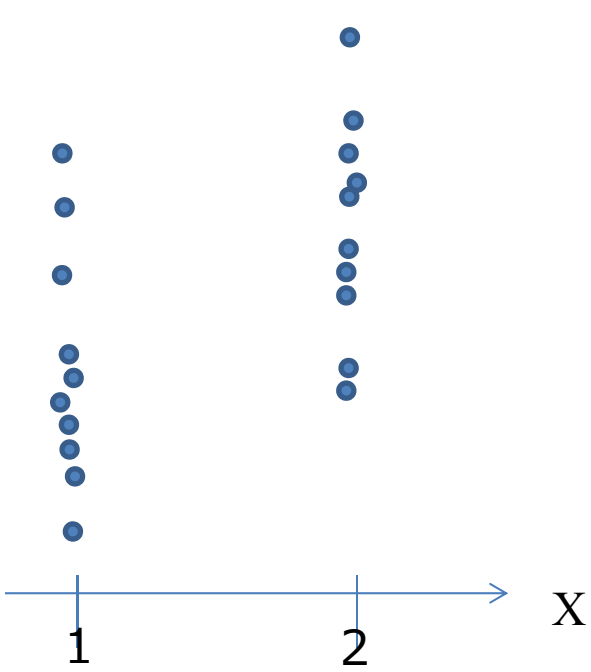
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.390012	3.100143	8.51	<.0001*
SepalWidth	0.6904897	0.089899	7.68	<.0001*

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

回帰分析と分散分析

Ｘが連続量 → 回帰分析　　Ｘが質的変数 → 分散分析



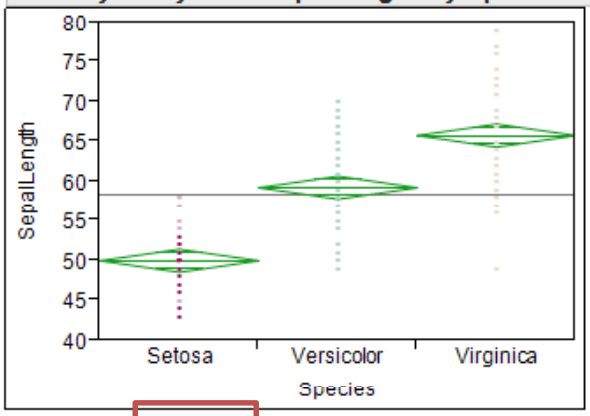


営利目的でのご利用はご遠慮ください

ICRweb http://www.icrweb.jp/

品種間でのがくの長さの違い

Oneway Analysis of SepalLength By Species



一元配置分散分析の結果

Oneway Anova

Summary of Fit

Rsquare	0.618706
Adj Rsquare	0.613518
Root Mean Square Error	5.147894
Mean of Response	58.43333
Observations (or Sum Wgts)	150

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Species	2	6321.213	3160.61	119.2645	<.0001*
Error	147	3895.620	26.50		
C. Total	149	10216.833			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Setosa	50	50.0600	0.72802	48.621	51.499
Versicolor	50	59.3600	0.72802	57.921	60.799
Virginica	50	65.8800	0.72802	64.441	67.319

Std Error uses a pooled estimate of error variance

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

理論展開の歴史

一般化線形モデル GLIN
Nelder&McCullugh(1983)
SAS/GENMOD

モデル拡張
非線型モデル
Bates&Watts(1988),
Draper&Smith SAS/NLIN

回帰分析
1877 F.Galton
regression, reversion
$y \sim f(x_1,\ldots,x_p)$

線形モデル

Probit解析
Finney(1960) SAS/PROBIT
logistic回帰
Cornfield(1957), Agresti(1990)
SAS/LOGISTIC
Cox回帰
Cox(1972) SAS/PHREG

位置の尺度の解析

Scheffe(1959)
Rao(1965,73)

成長曲線

バラツキの尺度の解析

混合モデル

分散分析 1920-30 R.A.Fisher
analysis of variance
responseの変動の分解

BMDPV,
SAS/GLM,
GLIM(GENSTAT)

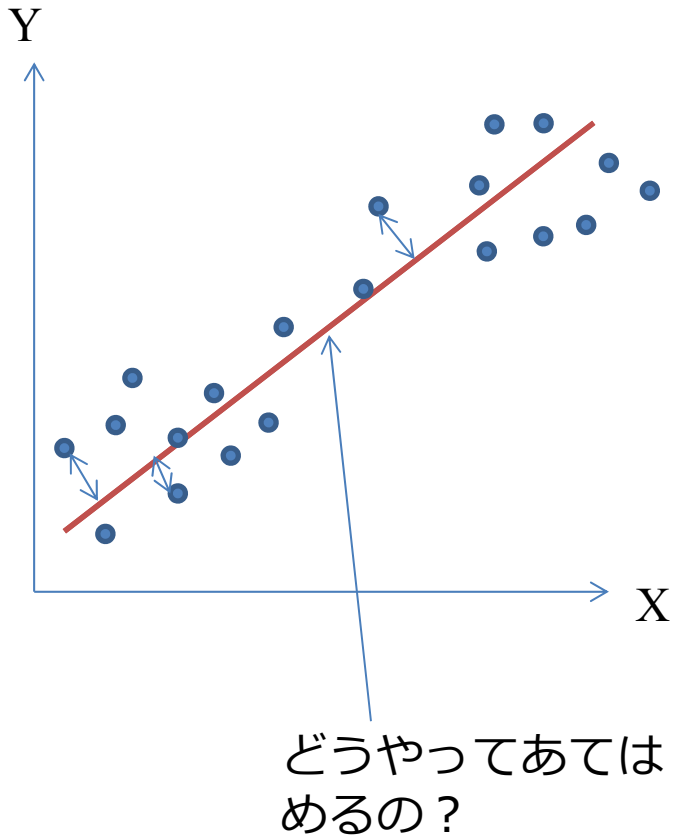
分散成分
Markov-Chainﾓﾝﾃｶﾙﾛ
variance component
Searle et al.(1992), Hocking(1988)
SAS/VARCOMP

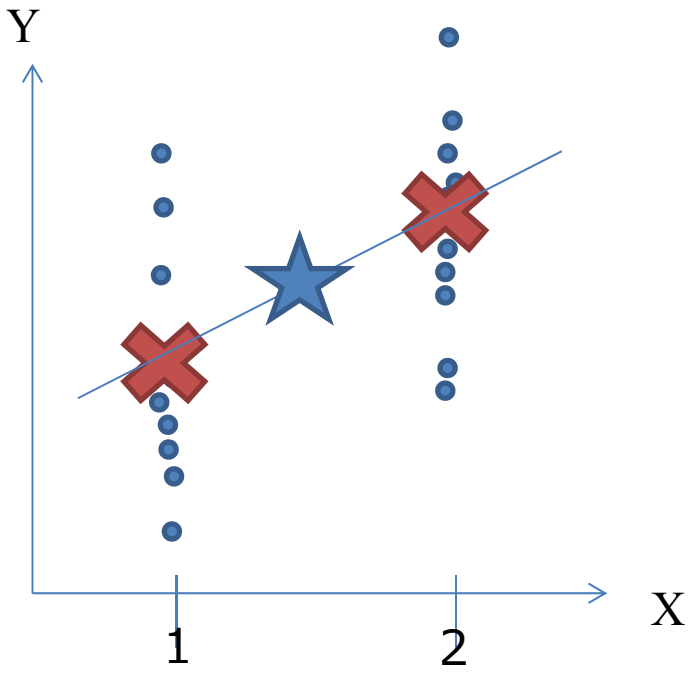
大橋靖雄先生講義資料を改変

営利目的でのご利用はご遠慮ください
ICRweb http://www.icrweb.jp/

回帰分析と分散分析

両方連続量 → 回帰分析　　片方質的変数 → 分散分析

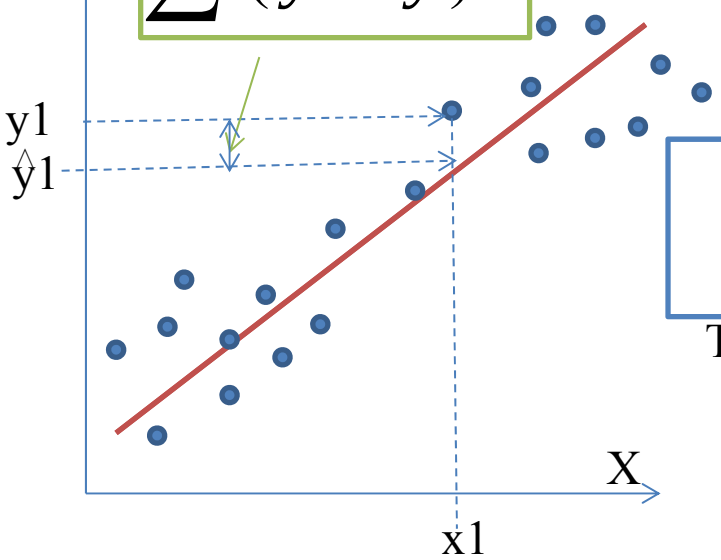




営利目的でのご利用はご遠慮ください　　ICRweb http://www.icrweb.jp/

回帰と最小2乗法

$$\sum (y-\hat{y})^2 \rightarrow \min$$



$$\underbrace{\sum (y-\bar{y})^2}_{\text{全平方和}} = \underbrace{\sum (y-\hat{y})^2}_{\text{残差平方和}} + \underbrace{\sum (\hat{y}-\bar{y})^2}_{\text{モデル平方和}}$$

	Total sum of squares		residual SS		model SS
	1	=	$(1-r^2)$	+	r^2
df	n-1	=	n-p-1		p

$$\sum (y-\bar{y})^2 / (n-1)$$

標本分散！！

営利目的でのご利用はご遠慮ください

ICRweb http://www.icrweb.jp/

（統計）モデル？

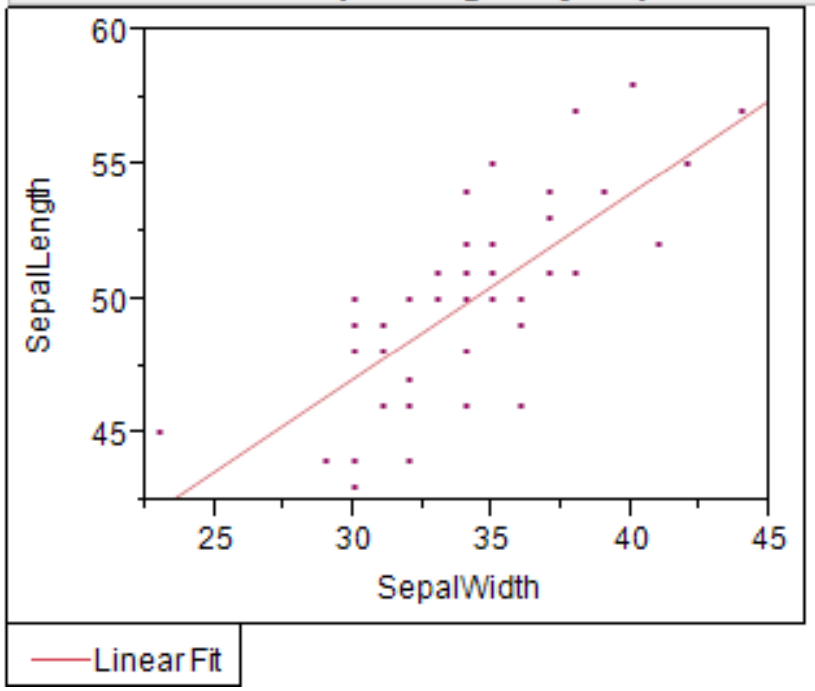
- 観測されたデータを説明できるようなモデルであれば、何でもよい。

- 線形？非線形？正規分布？二項分布？・・・

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

がくの幅と長さの関係
ー線形モデルのあてはめー

Bivariate Fit of SepalLength By SepalWidth Species=Setosa



Linear Fit

SepalLength = 26.390012 + 0.6904897*SepalWidth

Summary of Fit

RSquare	0.551376
RSquare Adj	0.542029
Root Mean Square Error	2.385422
Mean of Response	50.06
Observations (or Sum Wgts)	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	335.68848	335.688	58.9937
Error	48	273.13152	5.690	Prob > F
C. Total	49	608.82000		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.390012	3.100143	8.51	<.0001*
SepalWidth	0.6904897	0.089899	7.68	<.0001*

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

がくの幅と長さの関係

Linear Fit

SepalLength = 26.390012 + 0.6904897*SepalWidth

Summary of Fit

RSquare	0.551376
RSquare Adj	0.542029
Root Mean Square Error	2.385422
Mean of Response	50.06
Observations (or Sum Wgts)	50

R^2=
(モデル平方和）/（全平方和）
＝335.69/608.82≒0.55

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	
Model	1	335.68848	335.688	58.9937	モデル平方和
Error	48	273.13152	5.690	Prob > F	残差平方和
C. Total	49	608.82000		<.0001*	全平方和

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

分散分析のF検定

- 平均平方＝1パラメタあたりの平方和
  ＝平方和/自由度

$$\sum (y-\bar{y})^2 \Big/ (n-1)$$

- 帰無仮説：

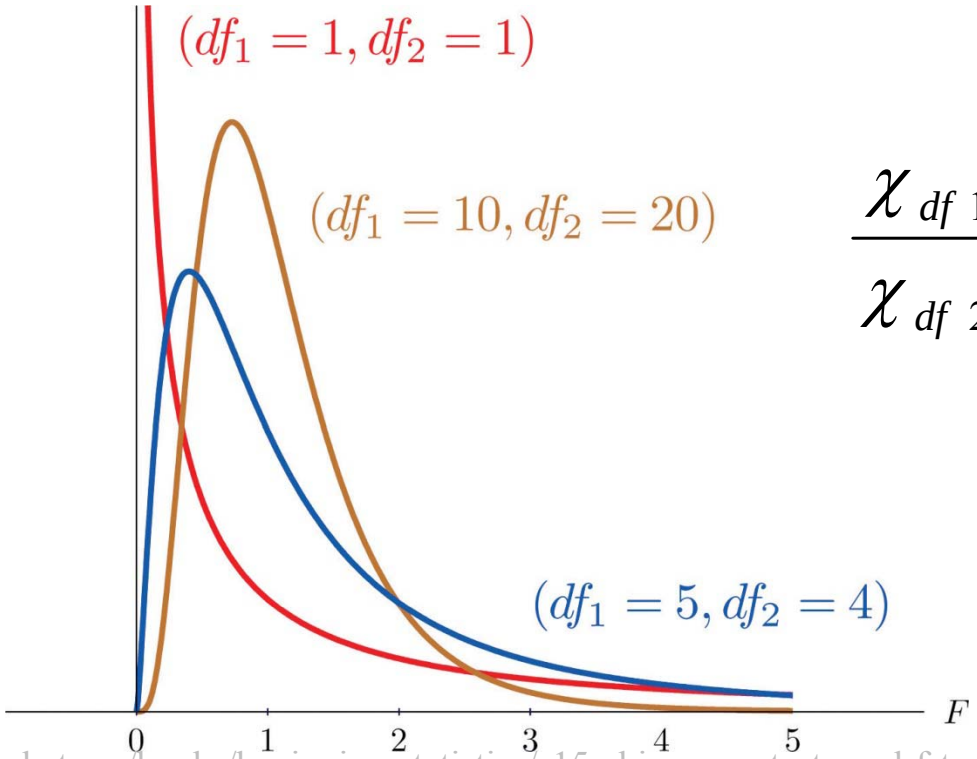
関心あるモデルの平均平方＝残差の平均平方

$$F = \frac{\text{関心あるモデル平方和／対応する自由度}}{\text{残差平方和／対応する自由度}}$$

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

F分布とF検定

- Ｘ２乗分布に従う確率変数の比は、Ｆ分布に従う



$$\frac{\chi_{df\,1}}{\chi_{df\,2}} \sim f_{(df\,1, df\,2)}$$

http://2012books.lardbucket.org/books/beginning-statistics/s15-chi-square-tests-and-f-tests.html

Ｆ検定とは、Ｆ分布に従う統計量を用いて行う統計学的検定の総称

営利目的でのご利用はご遠慮ください

がくの幅と長さの関係

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	
Model	1	335.68848	335.688	58.9937	モデル平方和
Error	48	273.13152	5.690	Prob > F	残差平方和
C. Total	49	608.82000		<.0001*	全平方和

335.688 ／1

$$F = \frac{\text{関心あるモデル平方和／対応する自由度}}{\text{残差平方和／対応する自由度}}$$

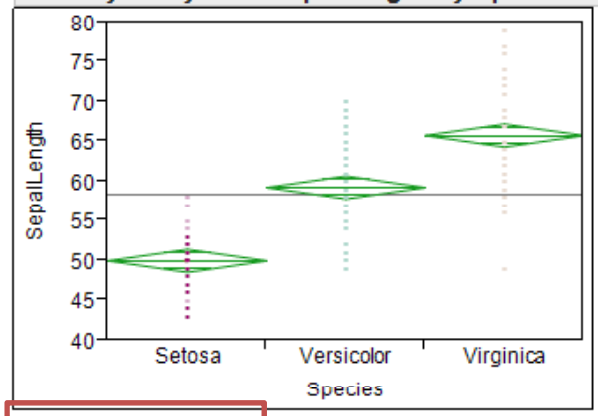
273.13 ／ 48

$$= 58.99$$

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

品種間でのがくの長さの違い

Oneway Analysis of SepalLength By Species



一元配置分散分析の結果

Oneway Anova

Summary of Fit

Rsquare	0.618706
Adj Rsquare	0.613518
Root Mean Square Error	5.147894
Mean of Response	58.43333
Observations (or Sum Wgts)	150

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Species	2	6321.213	3160.61	119.2645	<.0001*
Error	147	3895.620	26.50		
C. Total	149	10216.833			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Setosa	50	50.0600	0.72802	48.621	51.499
Versicolor	50	59.3600	0.72802	57.921	60.799
Virginica	50	65.8800	0.72802	64.441	67.319

Std Error uses a pooled estimate of error variance

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

品種間でのがくの長さの違い
一元配置分散分析

Oneway Anova

Summary of Fit

Rsquare	0.618706
Adj Rsquare	0.613518
Root Mean Square Error	5.147894
Mean of Response	58.43333
Observations (or Sum Wgts)	150

6321.213/10216.833

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Species	2	6321.213	3160.61	119.2645	<.0001*
Error	147	3895.620	26.50		
C. Total	149	10216.833			

3160.61/26.5

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

一元配置分散分析の場合のＦ検定

$$F = \frac{\text{関心あるモデル平方和／対応する自由度}}{\text{残差平方和／対応する自由度}}$$

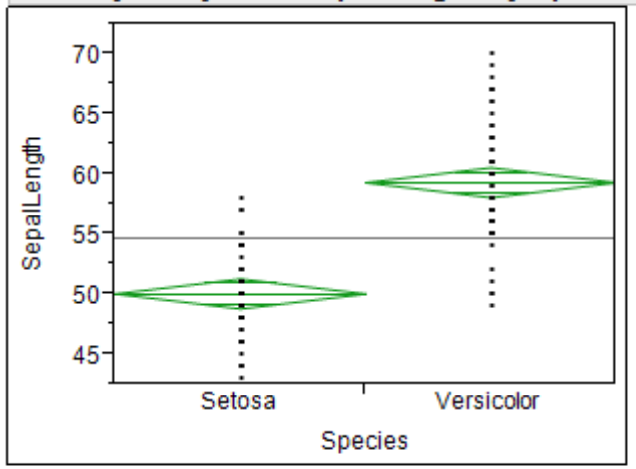
$$= \frac{\text{群間平方和（BSS: between sum of square)／対応する自由度}}{\text{群内平方和（WSS: within sum of square)／対応する自由度}}$$

$$\underbrace{\boxed{\sum (y_{ij} - \bar{y}_{..})^2}}_{\substack{\text{全平方和} \\ \text{TSS:} \\ \text{Total sum of squares}}} = \underbrace{\boxed{\sum\sum (y_{ij} - \bar{y}_{i.})^2}}_{\substack{\text{残差平方和} \\ \text{群内平方和} \\ \text{RSS/WSS}}} + \underbrace{\boxed{\sum\sum (\bar{y}_{i.} - \bar{y}_{..})^2}}_{\substack{\text{モデル平方和} \\ \text{群間平方和} \\ \text{MSS/BSS}}}$$

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

t検定と分散分析

Oneway Analysis of SepalLength By Species



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Species	1	2162.2500	2162.25	110.6912	<.0001*
Error	98	1914.3400	19.53		
C. Total	99	4076.5900			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Setosa	50	50.0600	0.62505	48.820	51.300
Versicolor	50	59.3600	0.62505	58.120	60.600

Std Error uses a pooled estimate of error variance

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
110.6912	1	86.538	<.0001*

t Test
10.5210

Oneway Anova

Summary of Fit

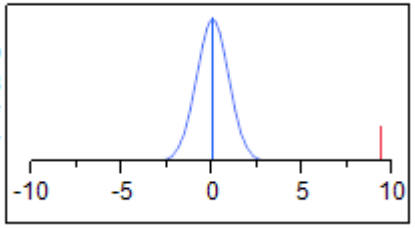
Rsquare	0.530407
Adj Rsquare	0.525615
Root Mean Square Error	4.419738
Mean of Response	54.71
Observations (or Sum Wgts)	100

t Test

Versicolor-Setosa

Assuming equal variances

Difference	9.3000	t Ratio	10.52099
Std Err Dif	0.8839	DF	98
Upper CL Dif	11.0542	Prob > |t|	<.0001*
Lower CL Dif	7.5458	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



2群比較の分散分析のＦ検定結果はt検定（ウェルチ）の結果と同じ

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

t検定

- 今1群ｎ人からなる群１の平均μ1と群２の平均μ2を比較したい
- 帰無仮説：μ1=μ2

$$t = \frac{(\mu 1 - \mu 2)}{\sqrt{\frac{s_1^{\ 2}}{n-1} + \frac{s_2^{\ 2}}{n-1}}}$$

（ただしウェルチ）

$$s_1^{\ 2} = \sum (y_j - \mu 1)^2$$

群内平方和！

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

2群比較のＦ検定

- 対立仮説：μ1≠μ2
- 帰無仮説：μ1＝μ2＝μ（全体平均）

$$F = \frac{\text{群間平方和（BSS: between sum of square)／対応する自由度}}{\text{群内平方和（WSS: within sum of square)／対応する自由度}}$$

$$= \frac{\sum(\mu 1-\mu)^2 + \sum(\mu 2-\mu)^2 / 1}{\sum(y_j-\mu 1)^2 + \sum(y_j-\mu 2)^2 / 2n-2}$$

$$= t^2$$

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

t検定と分散分析

- 2群比較の場合はt検定とＦ検定の結果はn→∞で同じ

$$
\begin{aligned}
\text{分子} &= \sum(\mu 1-\mu)^2+\sum(\mu 2-\mu)^2 \\
&= \sum_{i=1}^{n}(\mu 1-(\frac{\mu 1+\mu 2}{2}))^2+\sum_{i=1}^{n}(\mu 2-(\frac{\mu 1+\mu 2}{2}))^2 \\
&= \sum_{i=1}^{n}(\frac{\mu 1}{2}-\frac{\mu 2}{2})^2+\sum_{i=1}^{n}(\frac{\mu 2}{2}-\frac{\mu 1}{2})^2 \\
&= n((\frac{\mu 1}{2})^2+(\frac{\mu 1}{2})^2-2\frac{\mu 1*\mu 2}{2}+(\frac{\mu 2}{2})^2+(\frac{\mu 1}{2})^2-2\frac{\mu 1*\mu 2}{2}) \\
&= \frac{n}{2}(\mu 1^2-2\mu 1*\mu 2+\mu 2^2) \\
&= \frac{n}{2}(\mu 1-\mu 2)^2
\end{aligned}
$$

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

t検定で出てくるF検定？

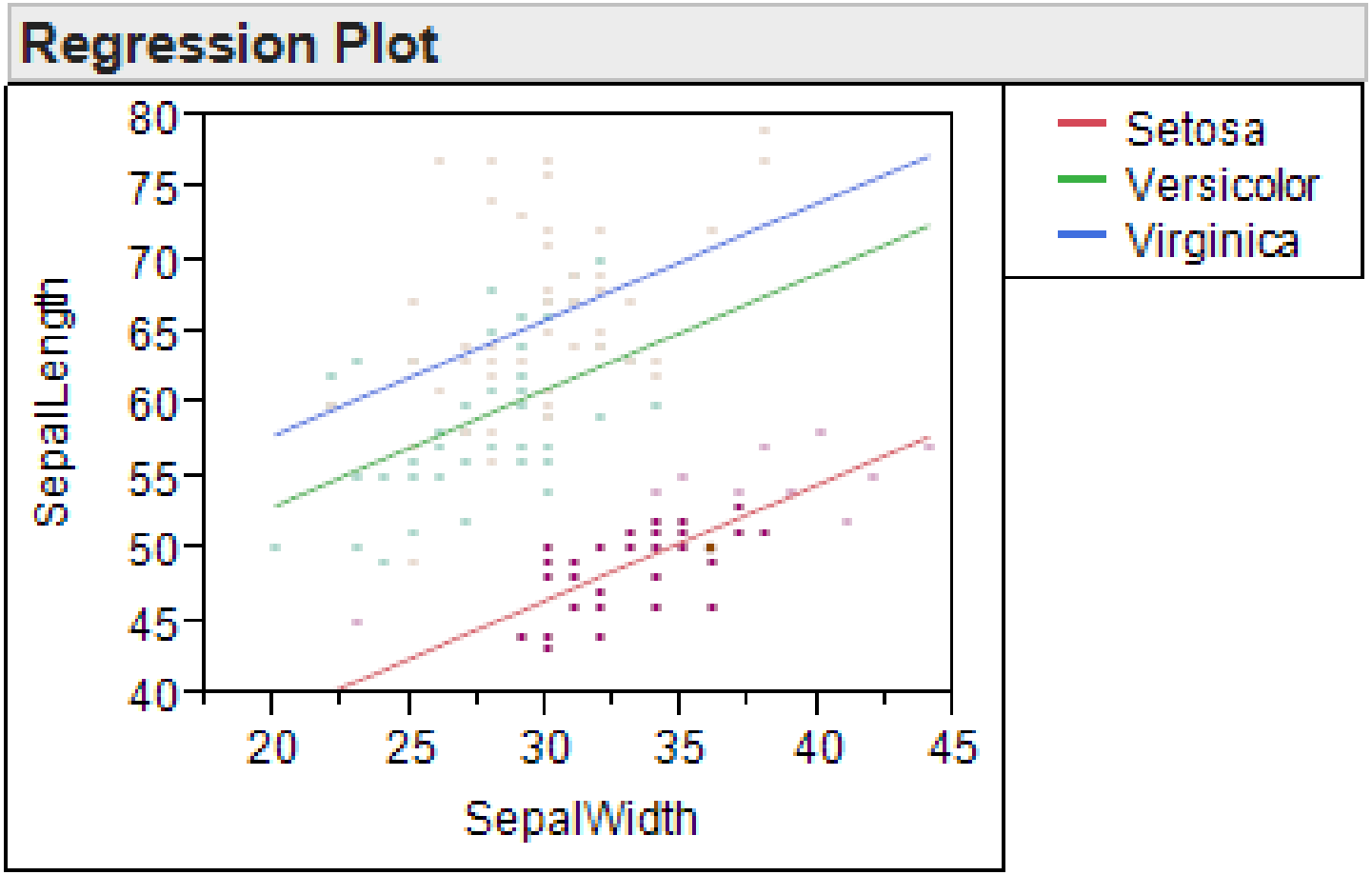
オリジナル t 検定は、2群に等分散を仮定しているので、ソフトウエアによっては以下の“不等分散性のF検定”を行ってくれる場合がある。

$$F = \frac{\text{群１の分散}}{\text{群２の分散}} \sim f_{(\text{群１の自由度},\ \text{群２の自由度})}$$

Ｆ検定とは、Ｆ分布に従う統計量を用いて行う統計学的検定の総称

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

品種間とがくの幅による
がくの長さの違い



営利目的でのご利用はご遠慮ください

ICRweb http://www.icrweb.jp/

品種間とがくの幅によるがくの長さの違い
共分散分析

Analysis of Variance

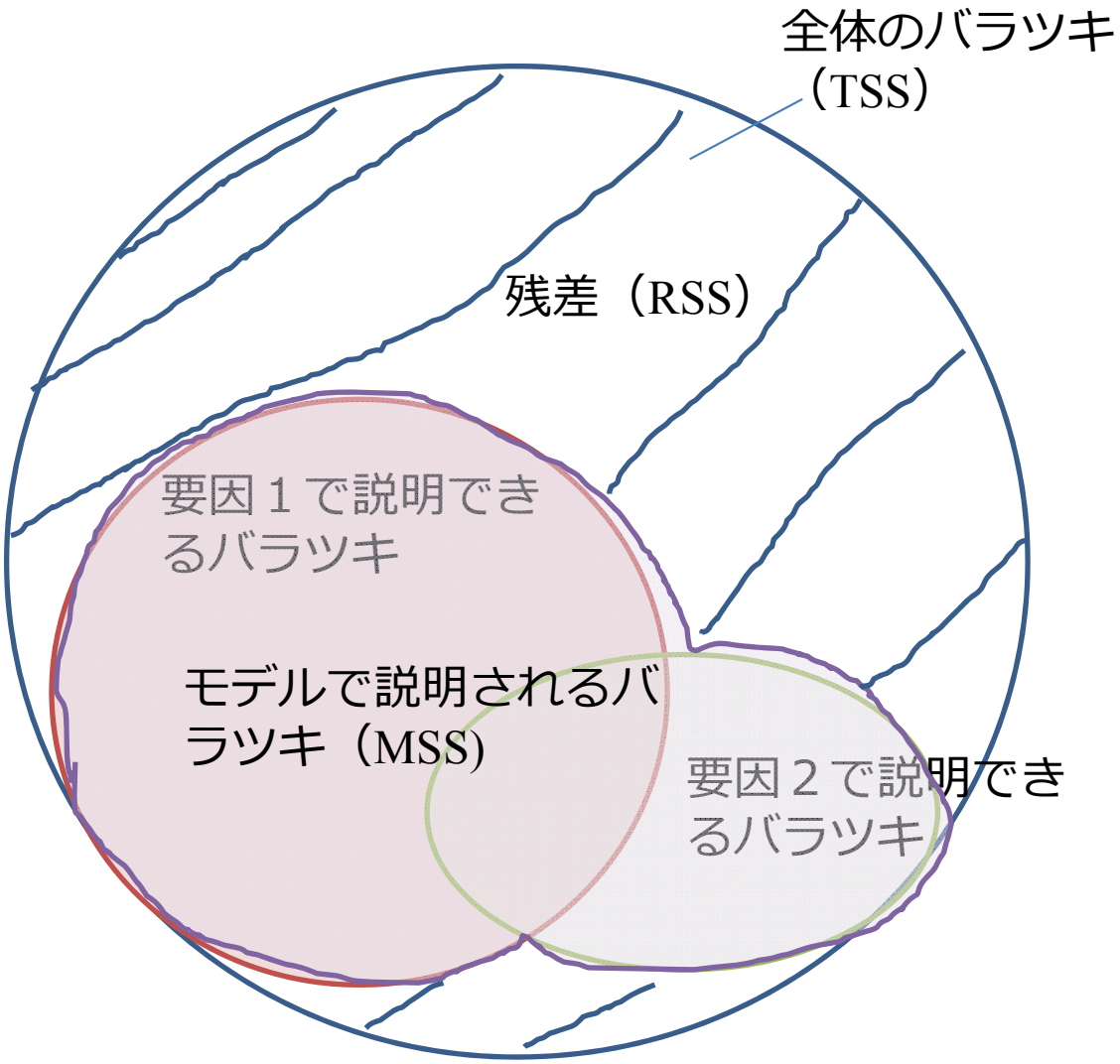
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	7416.467	2472.16	128.8884
Error	146	2800.366	19.18	Prob > F
C. Total	149	10216.833		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Species	2	2	7275.2431	189.6512	<.0001*
SepalWidth	1	1	1095.2535	57.1022	<.0001*

営利目的でのご利用はご遠慮ください ICRweb http://www.icrweb.jp/

平方和の分解



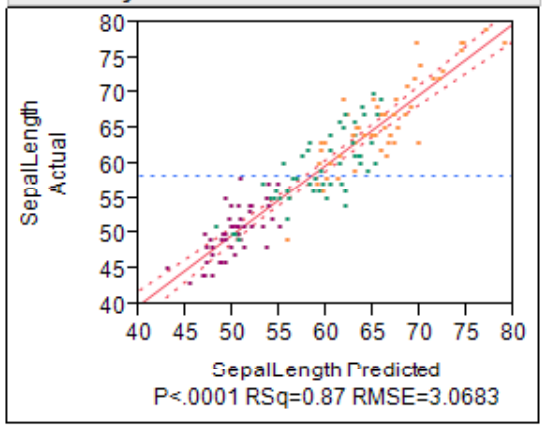
営利目的でのご利用はご遠慮ください
ICRweb http://www.icrweb.jp/

多変量モデル

Response SepalLength

Whole Model

Actual by Predicted Plot



Summary of Fit

RSquare	0.867312
RSquare Adj	0.862705
Root Mean Square Error	3.068261
Mean of Response	58.43333
Observations (or Sum Wgts)	150

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	8861.185	1772.24	188.2510
Error	144	1355.649	9.41	Prob > F
C. Total	149	10216.833		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	137	1341.6485	9.79305	4.8965
Pure Error	7	14.0000	2.00000	Prob > F
Total Error	144	1355.6485		0.0159*

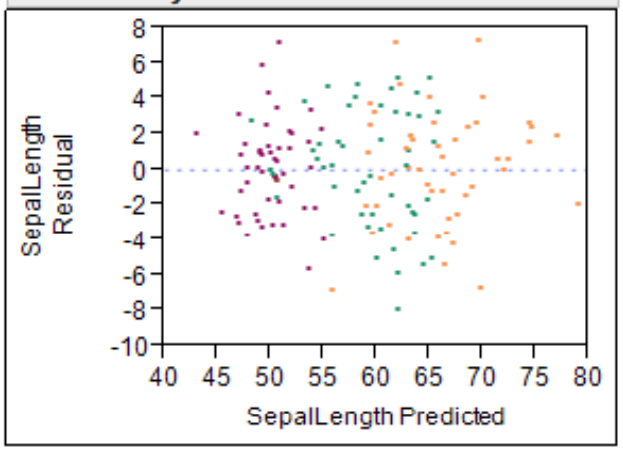
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	15.88913	2.716957	5.85	<.0001*
Species[Setosa]	5.8235326	1.897225	3.07	0.0026*
Species[Versicolor]	-1.412087	0.5991	-2.36	0.0198*
SepalWidth	0.4958889	0.08607	5.76	<.0001*
PetalLength	0.8292439	0.068528	12.10	<.0001*
PetalWidth	-0.315155	0.151196	-2.08	0.0389*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Species	2	2	88.8920	4.7212	0.0103*
SepalWidth	1	1	312.5004	33.1945	<.0001*
PetalLength	1	1	1378.5343	146.4310	<.0001*
PetalWidth	1	1	40.9029	4.3448	0.0389*

Residual by Predicted Plot



営利目的でのご利用はご遠慮ください
ICRweb http://www.icrweb.jp/

分散分析　まとめ

- 分散分析のF検定で有意差があった

＝なにかしらの統計モデルを仮定しない場合よりも仮定したモデルがよくデータを説明できた

- 多群の群ごとの比較を行いたい場合などは、多重性を調整するなどの、もう少し上級編が必要。

- バラツキ尺度の解析を行う場合でも、基本的な考え方は一緒

営利目的でのご利用はご遠慮ください

営利目的でのご利用はご遠慮ください

ICRweb http://www.icrweb.jp/