STA221

Neil Montgomery

Last edited: 2017-07-20 18:59

the sample correlation coefficient

Recall this expression that is used in the formula for b_1:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

the sample correlation coefficient

Recall this expression that is used in the formula for b_1:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

This is a (crude) measure of the linear association between dataset variables with names x and y.

the sample correlation coefficient

Recall this expression that is used in the formula for b_1:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

This is a (crude) measure of the linear association between dataset variables with names x and y.

It turns out to be a variation on something called a "sample covariance":

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

the sample correlation coefficient

The sample covariance s_{xy} depends on the unit of measurement for both variables, when all we care about is the strength of the relationship.

We can divide out the variation in both x and y to obtain what is called the *sample correlation coefficient:*

the sample correlation coefficient

The sample covariance s_{xy} depends on the unit of measurement for both variables, when all we care about is the strength of the relationship.

We can divide out the variation in both x and y to obtain what is called the *sample correlation coefficient:*

$$r = \frac{s_{xy}}{s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where s_x and s_y are the sample standard deviations of the x and y variables, and the final expression because the $n - 1$ cancels top and bottom.

the sample correlation coefficient

The sample covariance s_{xy} depends on the unit of measurement for both variables, when all we care about is the strength of the relationship.

We can divide out the variation in both x and y to obtain what is called the *sample correlation coefficient:*

$$r = \frac{s_{xy}}{s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where s_x and s_y are the sample standard deviations of the x and y variables, and the final expression because the $n - 1$ cancels top and bottom.

The sample mean estimates the mean... The sample variance estimates the variance... The sample correlation coefficient does in fact estimate a true, unknown "correlation coefficient", which is called ρ, but whose details we will not investigate, other than to point out that it is a number that assesses the strenght of the linear relationship between two distributions.

properties of the sample correlation coefficient

It is symmetric in x and y. There is not (necessarily) an "input" and an "output" variable.

properties of the sample correlation coefficient

It is symmetric in x and y. There is not (necessarily) an "input" and an "output" variable.

$$(r)^2 = R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

properties of the sample correlation coefficient

It is symmetric in x and y. There is not (necessarily) an "input" and an "output" variable.

$$(r)^2 = R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$-1 \leq r \leq 1$$

properties of the sample correlation coefficient

It is symmetric in x and y. There is not (necessarily) an "input" and an "output" variable.

$$(r)^2 = R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$-1 \leq r \leq 1$$

The sample correlation coefficient is only suitable when the relationship is linear, and is susceptible to all the same shortcomings as any regression model.

properties of the sample correlation coefficient

It is symmetric in x and y. There is not (necessarily) an "input" and an "output" variable.

$$(r)^2 = R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$-1 \leq r \leq 1$$

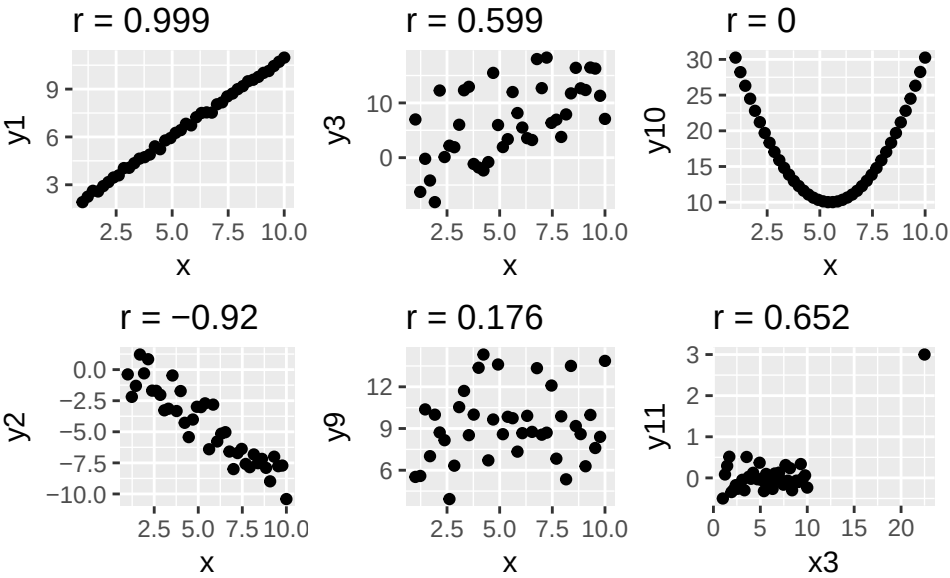
The sample correlation coefficient is only suitable when the relationship is linear, and is susceptible to all the same shortcomings as any regression model.

CORRECTED!

$$r = b_1 \sqrt{\frac{S_{yy}}{S_{xx}}}$$

where b_1 is the slope estimator with x is "input"...

examples


r = 0.999
y1
9
6
3
2.5 5.0 7.5 10.0
x
r = 0.599
y3
10
0
2.5 5.0 7.5 10.0
x
r = 0
y10
30
25
20
15
10
2.5 5.0 7.5 10.0
x
r = −0.92
y2
0.0
−2.5
−5.0
−7.5
−10.0
2.5 5.0 7.5 10.0
x
r = 0.176
y9
12
9
6
2.5 5.0 7.5 10.0
x
r = 0.652
y11
3
2
1
0
0 5 10 15 20
x3

inference for correlation coefficient

Since b_1 has a normal distribtion, it might not come as a surprise the r also has a normal distribution. In fact:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

inference for correlation coefficient

Since b_1 has a normal distribtion, it might not come as a surprise the r also has a normal distribution. In fact:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

So it is possible to do hypothesis testing for $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$.

inference for correlation coefficient

Since b_1 has a normal distribtion, it might not come as a surprise the r also has a normal distribution. In fact:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

So it is possible to do hypothesis testing for $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$.

(Note: confidence interval is also possible, but this is best left to the computer.)

bodyfat example

Recall the dataset:

```
## # A tibble: 250 × 15
##   `Pct BF`   Age Weight Height  Neck Chest Abdomen    waist   Hip
##      <dbl> <int>  <dbl>  <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl>
## 1     12.3    23 154.25  67.75  36.2  93.1    85.2 33.54331  94.5
## 2      6.1    22 173.25  72.25  38.5  93.6    83.0 32.67717  98.7
## 3     25.3    22 154.00  66.25  34.0  95.8    87.9 34.60630  99.2
## 4     10.4    26 184.75  72.25  37.4 101.8    86.4 34.01575 101.2
## 5     28.7    24 184.25  71.25  34.4  97.3   100.0 39.37008 101.9
## # ... with 245 more rows, and 6 more variables: Thigh <dbl>,
## #   Knee <dbl>, Ankle <dbl>, Bicep <dbl>, Forearm <dbl>, Wrist <dbl>
```

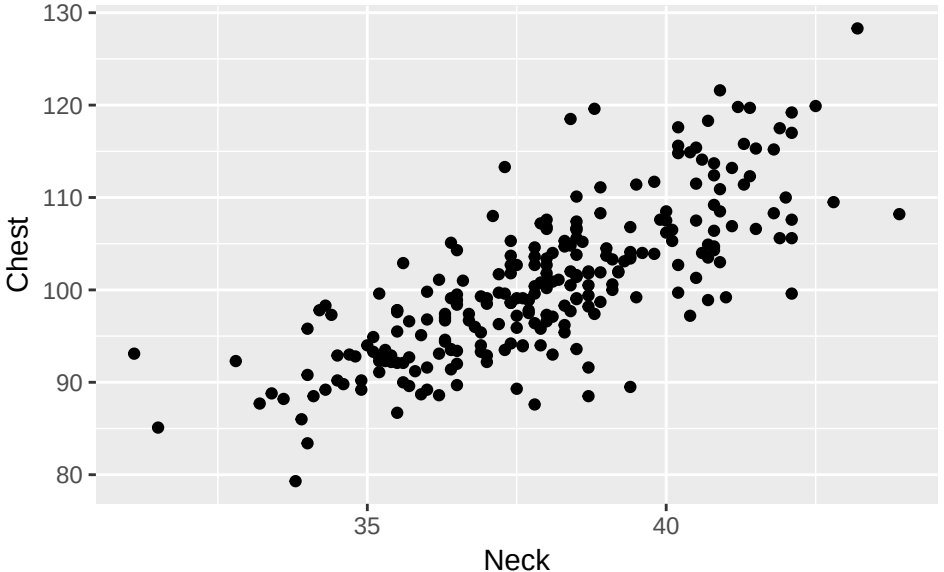
I wonder if the correlation between `Neck` and `Chest` circumferences is non-zero.

example - correlation matrix

A very useful information display is a "correlation matrix". Focus on the nine displayed variables, excluding Age:

```
##           Pct BF Weight  Height  Neck Chest Abdomen waist   Hip
## Pct BF    1.0000  0.617 -0.0294 0.489 0.701   0.824 0.824 0.633
## Weight    0.6173  1.000  0.5129 0.810 0.891   0.874 0.874 0.933
## Height   -0.0294  0.513  1.0000 0.325 0.224   0.187 0.187 0.397
## Neck      0.4885  0.810  0.3247 1.000 0.769   0.728 0.728 0.708
## Chest     0.7007  0.891  0.2236 0.769 1.000   0.910 0.910 0.825
## Abdomen   0.8237  0.874  0.1867 0.728 0.910   1.000 1.000 0.861
## waist     0.8237  0.874  0.1867 0.728 0.910   1.000 1.000 0.861
## Hip       0.6327  0.933  0.3967 0.708 0.825   0.861 0.861 1.000
```

Neck versus Chest



correlation analysis

```
##
##  Pearson's product-moment correlation
##
## data:  Neck and Chest
## t = 20, df = 200, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.713 0.815
## sample estimates:
##   cor
## 0.769
```

another example: Pct BF versus Height

Recall:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.5808    14.1540    1.81    0.072
## Height       -0.0932     0.2012   -0.46    0.644
##
## Residual standard error: 8.31 on 248 degrees of freedom
## Multiple R-squared:  0.000864,   Adjusted R-squared:  -0.00317
## F-statistic: 0.214 on 1 and 248 DF,  p-value: 0.644
```

compare p-value of 0.644 for $H_0 : \beta_1 = 0$

Now the correlation analysis:

```
##
##  Pearson's product-moment correlation
##
## data:  Pct BF and Height
## t = -0.463, df = 248, p-value = 0.644
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1528991  0.0950239
## sample estimates:
##        cor
## -0.0293896
```

Not a coincidence! The conclusion must be identical.

the analysis of designed experiments

Formal definitions: factor, level

A *factor* is a controllable experimental condition.

Formal definitions: factor, level

A *factor* is a controllable experimental condition.

A factor can take on two or more *levels*.

Formal definitions: factor, level

A *factor* is a controllable experimental condition.

A factor can take on two or more *levels*.

E.g., in a study of haul trucks "oil brand" could be a factor, with levels "Castrol", "Volvo", "Komatsu".

Formal definitions: factor, level

A *factor* is a controllable experimental condition.

A factor can take on two or more *levels*.

E.g., in a study of haul trucks "oil brand" could be a factor, with levels "Castrol", "Volvo", "Komatsu".

When experimental units are randomly assigned to levels of a factor and some output measure is observed, this is called a *designed experiment*. The formal model is typically written as (more on this later):

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Formal definitions: factor, level

A *factor* is a controllable experimental condition.

A factor can take on two or more *levels*.

E.g., in a study of haul trucks "oil brand" could be a factor, with levels "Castrol", "Volvo", "Komatsu".

When experimental units are randomly assigned to levels of a factor and some output measure is observed, this is called a *designed experiment*. The formal model is typically written as (more on this later):

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

You've seen the case of $i \in \{1, 2\}$—such an experiment would be analyzed using a two-sample t procedure.

Formal definitions: factor, level

A *factor* is a controllable experimental condition.

A factor can take on two or more *levels*.

E.g., in a study of haul trucks "oil brand" could be a factor, with levels "Castrol", "Volvo", "Komatsu".

When experimental units are randomly assigned to levels of a factor and some output measure is observed, this is called a *designed experiment*. The formal model is typically written as (more on this later):

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

You've seen the case of $i \in \{1, 2\}$—such an experiment would be analyzed using a two-sample t procedure.

In reality any dataset with one categorical "input" variable and one numerical "output" variable will be analysed the same as a formally designed experiment.

Typical dataset...

Truck.ID	Oil	Viscosity
HT 265	Volvo	25.5
HT 372	Castrol	25.7
HT 572	Komatsu	25.6
HT 908	Volvo	24.7
HT 201	Castrol	26.5
HT 898	Komatsu	25.4
HT 944	Volvo	24.4
HT 660	Castrol	22.8
HT 629	Komatsu	26.1
HT 61	Volvo	25.0
HT 205	Castrol	25.0
HT 176	Komatsu	25.9

One factor notation, models

"Balanced" case with equal sample size n for each of k levels for $N = nk$ total.

Levels:	1	2	...	i	...	k
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}
Sample average:	$\overline{y}_1$	$\overline{y}_2$	$\cdots$	$\overline{y}_i$	$\cdots$	$\overline{y}_k$

Grand overall average: $\overline{\overline{y}}$

Models:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2)$$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \sum \alpha_i = 0 \qquad \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2)$$

The main question

The main question is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus the negation (equivalently: all the $\alpha_i = 0$.)

In other words "is the variation among all the y_{ij} due to the factor variable, or just due to random chance?". The analysis even follows this logic.

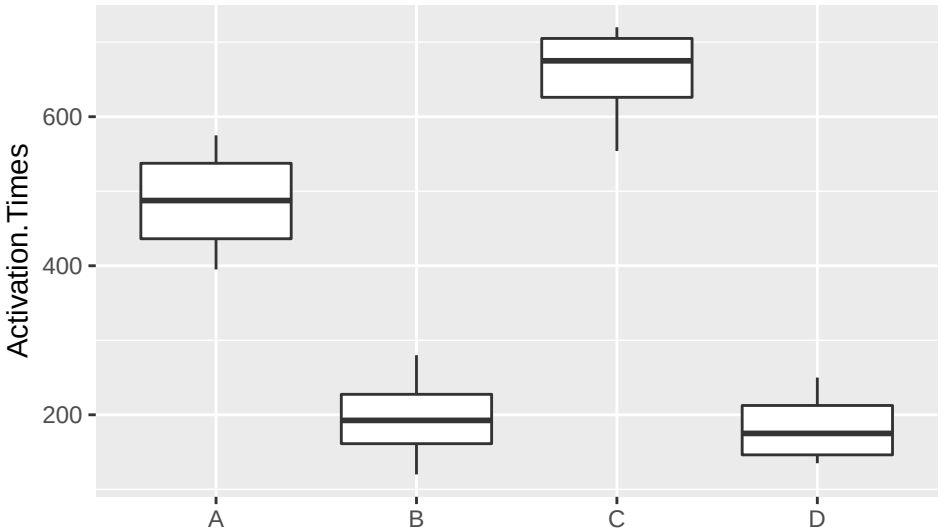
The variation among the y_{ij} is quantified as (as usual?):

$$(N-1) \cdot s_y^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \overline{\overline{y}}\right)^2$$

We will split this up into the "factor" part and the "random chance" part (like done in regression).

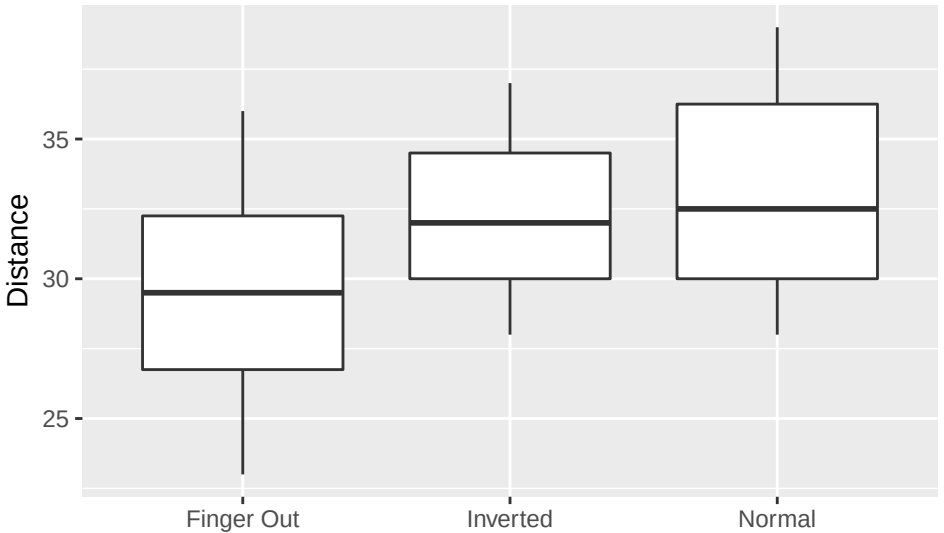
groups that are clearly different

From Q26.7 “Activating baking yeast”.



groups that aren't all that different

From Q26.8 "Frisbee throws".



some gory details

Build up from the inside out. For any i and j fixed:

$$(y_{ij} - \overline{\overline{y}})^2 = (y_{ij} - \bar{y}_i + \bar{y}_i - \overline{\overline{y}})^2$$

some gory details

Build up from the inside out. For any i and j fixed:

$$\begin{aligned}(y_{ij} - \bar{\bar{y}})^2 &= (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{\bar{y}})^2 \\ &= (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{\bar{y}})^2 + 2\,(y_{ij} - \bar{y}_i)\,(\bar{y}_i - \bar{\bar{y}})\end{aligned}$$

some gory details

Build up from the inside out. For any i and j fixed:

$$\begin{aligned}(y_{ij} - \bar{\bar{y}})^2 &= (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{\bar{y}})^2 \\ &= (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{\bar{y}})^2 + 2\,(y_{ij} - \bar{y}_i)\,(\bar{y}_i - \bar{\bar{y}})\end{aligned}$$

Next, for any fixed i, sum from $j = 1$ to n to get:

$$\sum_{j=1}^{n} (y_{ij} - \bar{\bar{y}})^2 = \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 + \sum_{j=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2 + 2\,(\bar{y}_i - \bar{\bar{y}}) \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)$$

some gory details

Build up from the inside out. For any i and j fixed:

$$\begin{aligned}(y_{ij} - \bar{\bar{y}})^2 &= (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{\bar{y}})^2 \\ &= (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{\bar{y}})^2 + 2\,(y_{ij} - \bar{y}_i)\,(\bar{y}_i - \bar{\bar{y}})\end{aligned}$$

Next, for any fixed i, sum from $j = 1$ to n to get:

$$\sum_{j=1}^{n} (y_{ij} - \bar{\bar{y}})^2 = \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 + \sum_{j=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2 + 2\,(\bar{y}_i - \bar{\bar{y}}) \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)$$

The term on the right hand side is always 0!

some gory details

Build up from the inside out. For any i and j fixed:

$$\begin{aligned}(y_{ij} - \bar{\bar{y}})^2 &= (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{\bar{y}})^2 \\ &= (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{\bar{y}})^2 + 2\,(y_{ij} - \bar{y}_i)\,(\bar{y}_i - \bar{\bar{y}})\end{aligned}$$

Next, for any fixed i, sum from $j = 1$ to n to get:

$$\sum_{j=1}^{n} (y_{ij} - \bar{\bar{y}})^2 = \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 + \sum_{j=1}^{n} (\bar{y}_i - \bar{\bar{y}})^2 + 2\,(\bar{y}_i - \bar{\bar{y}}) \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)$$

The term on the right hand side is always 0!

Finally, sum from $i = 1$ to k and rearrange:

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{k} n\,(\bar{y}_i - \bar{\bar{y}})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2=\sum_{i=1}^{k}n\left(\bar{y}_i-\overline{\overline{y}}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$

$$\begin{array}{ccccc} SS_{Total} & = & SS_T & + & SS_E \\ N-1 & = & k-1 & + & N-k \end{array}$$

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i-\overline{\overline{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$

$$\begin{array}{ccccc} SS_{Total} & = & SS_T & + & SS_E \\ N-1 & = & k-1 & + & N-k \end{array}$$

Holding SS_{Total} fixed, what would it mean for one or the other of SS_T and SS_E to be large?

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i-\overline{\overline{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$

$$\begin{array}{ccccc} SS_{Total} & = & SS_T & + & SS_E \\ N-1 & = & k-1 & + & N-k \end{array}$$

Holding SS_{Total} fixed, what would it mean for one or the other of SS_T and SS_E to be large?

It turns out we'll look at a ratio of SS_T and SS_E to make our final decision.

more details

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{\overline{y}}\right)^2 = \sum_{i=1}^{k} n\left(\bar{y}_i-\overline{\overline{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_i\right)^2$$

$$\begin{array}{ccccc} SS_{Total} & = & SS_T & + & SS_E \\ N-1 & = & k-1 & + & N-k \end{array}$$

Holding SS_{Total} fixed, what would it mean for one or the other of SS_T and SS_E to be large?

It turns out we'll look at a ratio of SS_T and SS_E to make our final decision.

From which family of distributions will SS_T and SS_E come from?

the F distributions

$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SS_E}{N-k}$$

the F distributions

$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SS_E}{N-k}$$

These are called "mean squares", and the ratio of mean squares will follow what is called an F distribution, with $k-1$ and $N-k$ "degrees of freedom".

the F distributions

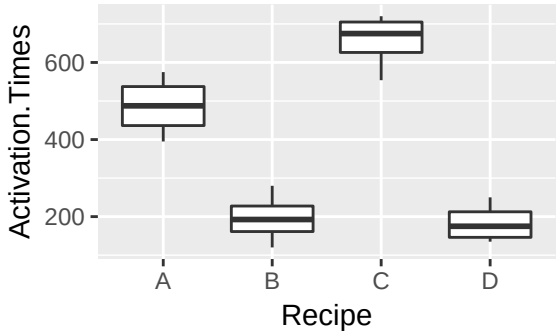
$$MS_T = \frac{SS_T}{k-1} \qquad \text{and} \qquad MS_E = \frac{SS_E}{N-k}$$

These are called "mean squares", and the ratio of mean squares will follow what is called an F distribution, with $k-1$ and $N-k$ "degrees of freedom".

When the null hypothesis is true, $\frac{MS_T}{MS_E}$ lives near 1, and large values of this ratio give small p-values.

putting it all together

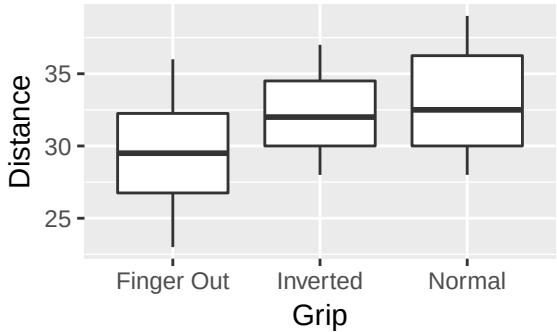
All this information is concisely displayed in what is called the "analysis of variance" table (or ANOVA table, or AOV table). Here's the table for the Yeast example:




```
##             Df Sum Sq Mean Sq F value     Pr(>F)
## Recipe       3 638968  212989    44.7 0.00000086
## Residuals   12  57128    4761
```

putting it all together

And for the "probably not different" Frisbee example:




```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Grip          2   58.6    29.3    2.05   0.15
## Residuals    21  300.8    14.3
```

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>					
Residuals					

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$				
Residuals	$N-k$				

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T			
Residuals	$N-k$	SS_E			

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$		
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	$P(F_{k-1,N-k} \geq F_{obs})$
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	$P(F_{k-1,N-k} \geq F_{obs})$
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

For example (from 25.13 "Hearing"). Four different word lists were compared for ease of hearing with background noise. 96 people were divided into four groups and the number out of 24 words understood was recorded.

ANOVA table—formula version

Not explicitly appearing on the R output is $SS_{Total} = SS_T + SS_E$ and $N - 1 = k - 1 + N - k$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<var_name>	$k-1$	SS_T	$MS_T = \frac{SS_T}{(k-1)}$	$F_{obs} = \frac{MS_T}{MS_E}$	$P(F_{k-1,N-k} \geq F_{obs})$
Residuals	$N-k$	SS_E	$MS_E = \frac{SS_E}{(N-k)}$		

For example (from 25.13 "Hearing"). Four different word lists were compared for ease of hearing with background noise. 96 people were divided into four groups and the number out of 24 words understood was recorded.

The sample variance for all 96 people was 70.090789. The mean squared error was 62.371. Is there a difference between the four word groups?

hand calculation example - hearing

The sample variance for all 96 people was 70.090789. The mean squared error was 62.371. Is there a difference between the four word groups?

hand calculation example - hearing

The sample variance for all 96 people was 70.090789. The mean squared error was 62.371. Is there a difference between the four word groups?

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
List	3	920.4583	306.819	4.9192	0.0033
Error	92	5738.1667	62.371		
Total	95	6658.6250			

the t - F connection - I

Any time you've done a t test, you could have done a (slightly inferior) F test.

the t - F connection - I

Any time you've done a t test, you could have done a (slightly inferior) F test.

That's because the square of anything with a t_ν distribution always has an $F_{1,\nu}$ distribution.

the t - F connection - I

Any time you've done a t test, you could have done a (slightly inferior) F test.

That's because the square of anything with a t_ν distribution always has an $F_{1,\nu}$ distribution.

For example, consider the two-sample t test (equal variance version using pooled variance s_p^2 - section 21.3 of the text).

Q21.20 "Hard Water" Mortality rates per county in 61 counties in England and Wales classified as "North" and "South" of Derby. Is there a difference in mean mortality rate? Here's the R output:

```
##
##  Two Sample t-test
##
## data:  Mortality by Derby
## t = 6.531, df = 59, p-value = 0.0000000167
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

mortality data via F test

From above:

`t = 6.5312, df = 59, p-value = 0.00000001673`

The `R` ANOVA output:

```
##              Df  Sum Sq Mean Sq F value      Pr(>F)
## Derby         1  886712  886712    42.7 0.000000017
## Residuals    59 1226462   20787
```

mortality data via F test

From above:

`t = 6.5312, df = 59, p-value = 0.00000001673`

The `R` ANOVA output:

```
##              Df  Sum Sq Mean Sq F value      Pr(>F)
## Derby         1  886712  886712    42.7 0.000000017
## Residuals    59 1226462   20787
```

Also, $6.5312^2 = 42.656573$.

the t - F connection - II

Recall from the `Bodyfat` example from regression:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.5808    14.1540    1.81    0.072
## Height       -0.0932     0.2012   -0.46    0.644
##
## Residual standard error: 8.31 on 248 degrees of freedom
## Multiple R-squared:  0.000864,   Adjusted R-squared:  -0.00317
## F-statistic: 0.214 on 1 and 248 DF,  p-value: 0.644
```

the t - F connection - II

Recall from the `Bodyfat` example from regression:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.5808    14.1540    1.81    0.072
## Height       -0.0932     0.2012   -0.46    0.644
##
## Residual standard error: 8.31 on 248 degrees of freedom
## Multiple R-squared:  0.000864,   Adjusted R-squared:  -0.00317
## F-statistic: 0.214 on 1 and 248 DF,  p-value: 0.644
```

Again, $t^2 = F$ and the p-values are identical.

the t - F connection - II

Recall from the `Bodyfat` example from regression:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.5808    14.1540    1.81    0.072
## Height       -0.0932     0.2012   -0.46    0.644
##
## Residual standard error: 8.31 on 248 degrees of freedom
## Multiple R-squared:  0.000864,   Adjusted R-squared:  -0.00317
## F-statistic: 0.214 on 1 and 248 DF,  p-value: 0.644
```

Again, $t^2 = F$ and the p-values are identical.

The practical downside of using F is that you lose information about the sign.

ANOVA model and calculations requirement

Look at the model again:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma)$$

First, the errors are supposed to be independent, which would exclude experiments such as giving the same person different treatments over time, etc.

ANOVA model and calculations requirement

Look at the model again:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma)$$

First, the errors are supposed to be independent, which would exclude experiments such as giving the same person different treatments over time, etc.

The main things to verify:

1. Do the groups come from a distribution with the same variance? (fatal if no)

ANOVA model and calculations requirement

Look at the model again:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma)$$

First, the errors are supposed to be independent, which would exclude experiments such as giving the same person different treatments over time, etc.

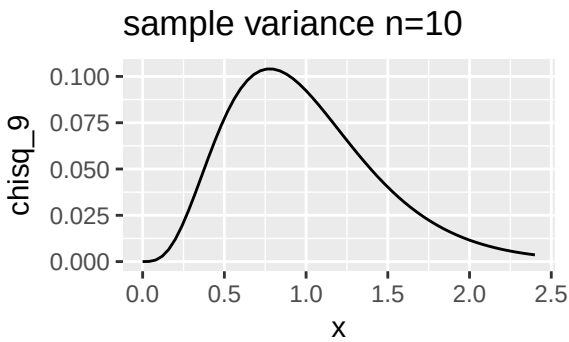
The main things to verify:

1. Do the groups come from a distribution with the same variance? (fatal if no)
2. Do the groups come from normal distributions? (OK if sample size is large enough)

Formal test for equality of variances

For the equal variance assumptions, the book says to look at plots. Other books gives a variety of heuristics. These suggestsions tend to be wildly conservative.

The problem is twofold. Within-group sample sizes tend to be small. And the sample variance itself has a very large variance.



Levene's test

Not in the book! And must be done on a computer.

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

versus at least two are unequal.

Levene's test

Not in the book! And must be done on a computer.

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

versus at least two are unequal.

The form of the test is exactly an ANOVA, but not on the original $y_i j$. Instead, it is on the *absolute differences from the group medians*:

$$Z_{ij} = |y_{ij} - \tilde{y}_i|$$

where $\tilde{y}_i$ is the sample median of the i^{th} group.

Levene's test

Not in the book! And must be done on a computer.

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2$$

versus at least two are unequal.

The form of the test is exactly an ANOVA, but not on the original y_ij. Instead, it is on the *absolute differences from the group medians*:

$$Z_{ij} = |y_{ij} - \tilde{y}_i|$$

where $\tilde{y}_i$ is the sample median of the i^{th} group.

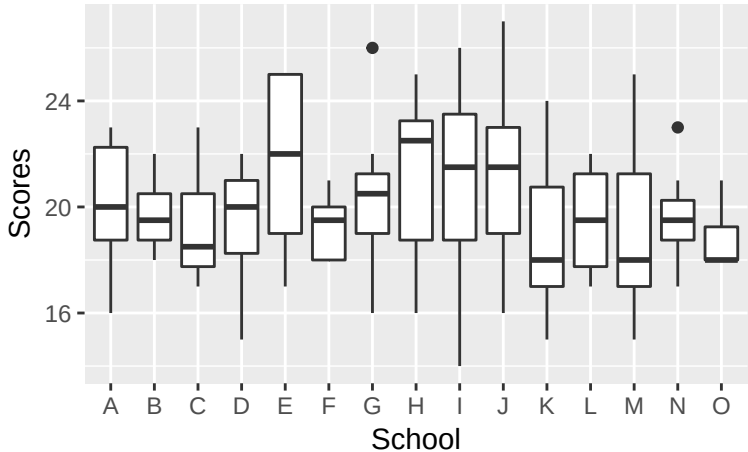
Plugging the Z_{ij} into the ANOVA formulae gives an approximate $F_{k-1,N-k}$ distribution.

Levene's test example - yeast

```
## # A tibble: 2 × 3
##      Df `F value` `Pr(>F)`
## * <int>     <dbl>    <dbl>
## 1     3  0.291724 0.830601
## 2    12        NA       NA
```

tougher example

From textbook question 25.18 “School System”. 15 schools selected. 8 students per school.



tougher example - Levene

```
## # A tibble: 2 × 3
##      Df `F value` `Pr(>F)`
## * <int>     <dbl>    <dbl>
## 1    14   1.26417 0.242227
## 2   105        NA       NA
```

normality assumption

Technically, all the groups have to be normal. But the samples sizes are usually too small.

normality assumption

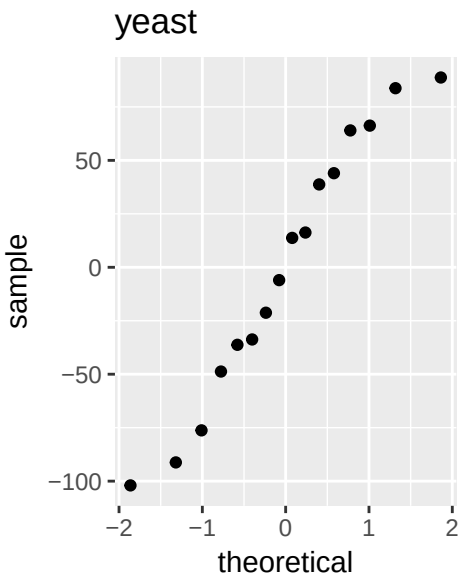
Technically, all the groups have to be normal. But the samples sizes are usually too small.

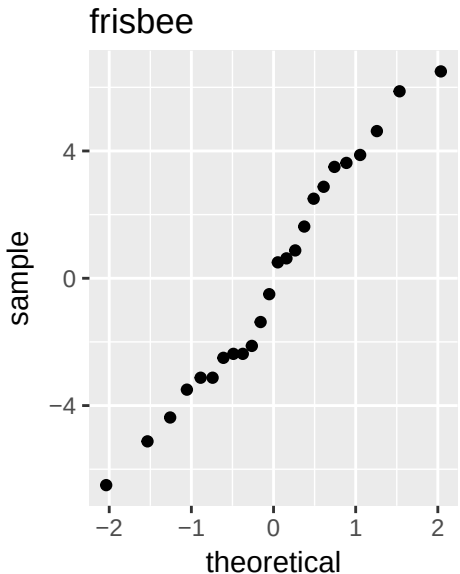
If the equal variance assumption has been satisfied (do that first), then the method is to pool all the *residuals* together:

$$y_{ij} - \overline{y}_i$$

and look at a normal quantile plot.

normal assumption verification examples





pairwise comparisons

The ANOVA F-test is an "omnibus" test—it tells you if there are *any* differences, without giving information about where the differences might be.

pairwise comparisons

The ANOVA F-test is an "omnibus" test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

pairwise comparisons

The ANOVA F-test is an "omnibus" test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

Sometimes one or more *pairwise* differences might be conceived of *in advance of collecting the data.*

pairwise comparisons

The ANOVA F-test is an "omnibus" test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

Sometimes one or more *pairwise* differences might be conceived of *in advance of collecting the data.*

Or, there may be some differences between groups that are noticed after collecting data, which is gets us into dangerous territory!

pairwise comparisons

The ANOVA F-test is an "omnibus" test—it tells you if there are *any* differences, without giving information about where the differences might be.

If the ANOVA F-test is large, there are no differences between groups of any kind, in which case what follows does not apply.

Sometimes one or more *pairwise* differences might be conceived of *in advance of collecting the data.*

Or, there may be some differences between groups that are noticed after collecting data, which is gets us into dangerous territory!

The approach in any case will be to perform multiple pooled two-sample t procedures, using the overall MSE in place of the usual pooled variance:

$$\frac{\overline{y}_i - \overline{y}_j}{\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim t_{N-k}$$

The usual technique is to produce confidence intervals for each desired pair. But at what confidence level? The usual 95% level leads to a problem...