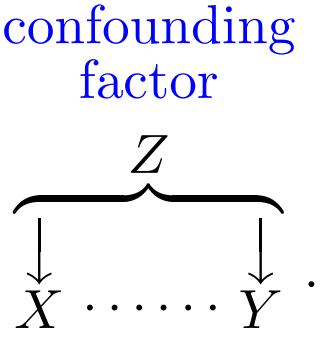
Association

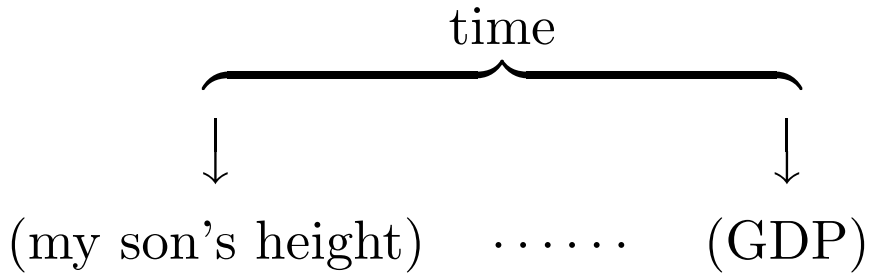
Observe $X \cdots\cdots Y$, but could be

$X \longrightarrow Y$ or $X \longleftarrow Y$ or



.

Silly Example



But absurd to conclude my son's height causes GDP growth.

Causation

Ideally,

$$\underbrace{\text{all else being equal}}_{(a)}, \quad \underbrace{\text{change } X \text{ and observe changes in } Y}_{(b)}.$$

(a) Pretty much impossible to implement.

(b) What kind of changes?
Often probabilistic rather than deterministic.

Key Device Randomization.

Randomization

Treatment (e.g., new drug)	Control (e.g., placebo)
$y_{1,t}$	$y_{1,c}$
$y_{2,t}$	$y_{2,c}$
$\vdots$	$\vdots$
$y_{n_t,t}$	$y_{n_c,c}$

$\Uparrow$ randomize $\Uparrow$

key distinction between experimental & observational studies

Observational Data

	Smoke (t)	Don't Smoke (c)
Cancer (1)	$n_{1,t}$	$n_{1,c}$
No Cancer (0)	$n_{0,t}$	$n_{0,c}$

$\Uparrow$ $\Uparrow$

cannot randomize

Possible Confounders Life style (drug use, diet, exercise)? Genetic?

Berkeley Admission 1973

	Men		Women	
	Applicants	% Admitted	Applicants	% Admitted
Total	2590	**46**	1835	30

Berkeley Admission 1973

	Men		Women	
Dept	Applicants	% Admitted	Applicants	% Admitted
A	825	62	108	**82**
B	560	63	25	**68**
C	325	37	593	34
D	417	33	375	**35**
E	191	28	393	24
F	272	6	341	**7**
Total	2590	**46**	1835	30

phenomenon referred to as "Simpson's Paradox"

Blocking

	Treatment (e.g., new drug)	Control (e.g., placebo)
elderly, diabetic	...	...
elderly, non-diabetic	...	...
young, diabetic	...	...
young, non-diabetic	...	...

$\Uparrow$ blocks

$\Uparrow$ $\Uparrow$ randomize

Go to Cartoon

One-Sample T-Test: Review

$$X_1, ..., X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$$

$$H_0 : \mu = 0$$

target	μ
estimator	$\bar{X}$
variance (of estimator)	σ^2/n

Under H_0,

$$\frac{\bar{X}}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0,1) \quad \Rightarrow \quad \frac{\bar{X}}{\sqrt{S^2/n}} \sim t_{(n-1)}$$

where

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Two-Sample T-Test

$$\{X_1, ..., X_n \overset{iid}{\sim} \mathrm{N}(\mu_1, \sigma^2)\} \quad \perp\!\!\!\perp \quad \{Y_1, ..., Y_m \overset{iid}{\sim} \mathrm{N}(\mu_2, \sigma^2)\}$$

$$H_0 : \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0$$

target	$\mu_1 - \mu_2$
estimator	$\bar{X} - \bar{Y}$
variance (of estimator)	$\sigma^2/n + \sigma^2/m$

Under H_0,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim \mathrm{N}(0, 1) \quad \Rightarrow \quad \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2/n + S_p^2/m}} \sim t_{(n+m-2)}$$

where

$$S_p^2 = \frac{1}{n+m-2} \left[\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{j=1}^{m} (Y_j - \bar{Y})^2 \right].$$

Back from Cartoon

Completely Randomized Design

General

Treatment Groups			
1	2	...	T
$y_{1,1}$	$y_{1,2}$	$\cdots$	$y_{1,T}$
$y_{2,1}$	$y_{2,2}$	$\cdots$	$y_{2,T}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$y_{n_1,1}$	$y_{n_2,2}$	$\cdots$	$y_{n_T,T}$

↑

will often suppress comma ","
y_{it} rather than $y_{i,t}$

Example

$T = 2$	
Gas A	Gas B
25.4	29.3
30.2	23.6
$\vdots$	$\vdots$
19.8	24.7

↑

y_{it}=mileage from different cabs

One-Way ANOVA

Model

$$y_{it} = \mu + \tau_t + \varepsilon_{it}, \quad \sum_t n_t \tau_t = 0 \quad (\text{why})$$

Null Hypothesis

$$H_0 : \tau_1 = \tau_2 = ... = \tau_T = 0$$

Exercise

$$\widehat{\mu} = \bar{y}_{\cdot\cdot} \qquad \text{and} \qquad \widehat{\tau}_t = \bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot}$$

One-Way ANOVA

Decomposition

$$\underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n_t}(y_{it}-\bar{y}_{\cdot\cdot})^2}_{\text{SS}_{total}} = \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n_t}[y_{it}-(\bar{y}_{\cdot t}-\bar{y}_{\cdot\cdot})-\bar{y}_{\cdot\cdot}]^2}_{\text{SS}_{err}}$$

$$+\underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n_t}(\bar{y}_{\cdot t}-\bar{y}_{\cdot\cdot})^2}_{\text{SS}_{treat}}$$

F-Test

$$\frac{\text{SS}_{treat}/(T-1)}{\text{SS}_{err}/(N-T)} \overset{H_0}{\sim} F_{(T-1,N-T)}, \qquad N=\sum_{t=1}^{T} n_t$$

Remark $\dim(M_A) = 1+(T-1)$, $\dim(M_0)=1$.

Nothing Special

Start with

$$\underbrace{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2}_{\text{denominator of } F} = \underbrace{\sum_t \sum_i [y_{it} - \overset{\widehat{\mu}}{\downarrow}\bar{y}_{\cdot\cdot} - \overbrace{(\bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot})}^{\widehat{\tau}_t}]^2}_{\text{SS}_{err}}$$

and

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_0\|^2 = \sum_t \sum_i (y_{it} - \overset{\widehat{\mu}}{\downarrow}\bar{y}_{\cdot\cdot})^2,$$

can show (exercise)

$$\underbrace{\overbrace{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_0\|^2}^{\text{SS}_{total}} - \overbrace{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2}^{\text{SS}_{within}}}_{\text{numerator of } F} = \underbrace{\overbrace{\sum_t \sum_i (\bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot})^2}^{\text{SS}_{between}}}_{\text{SS}_{treat}}.$$

Special Case: $T = 2$

Exercise Show that, when $T = 2$, the one-way ANOVA F-test is the same as the two-sample t-test:

$$\frac{\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_2-2},$$

where

$$s_p^2 = \frac{\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_{\cdot 1})^2 + \sum_{i=1}^{n_2} (y_{i2} - \bar{y}_{\cdot 2})^2}{n_1 + n_2 - 2}.$$

Hint Specifically, show that

$$(\text{LHS})^2 = \frac{\text{SS}_{treat}/(T-1)}{\text{SS}_{err}/(N-T)},$$

where now $T = 2$ and $N = n_1 + n_2$.

Randomized Block Design

General

	Treatment Groups			
	1	2	...	T
Block 1	y_{11}	y_{12}	$\cdots$	y_{1T}
Block 2	y_{21}	y_{22}	$\cdots$	y_{2T}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Block B	y_{B1}	y_{B2}	$\cdots$	y_{BT}

Example

	$T = 2$	
	Gas A	Gas B
Driver 01	20.0	20.4
Driver 02	30.1	29.6
$\vdots$	$\vdots$	$\vdots$
Driver 10	25.2	26.0

Two-Way ANOVA

Model

$$y_{bt} = \mu + \tau_t + \beta_b + \varepsilon_{bt}, \quad \sum_t \tau_t = 0, \quad \sum_b \beta_b = 0$$

Null Hypothesis

$$H_0 : \tau_1 = \tau_2 = ... = \tau_T = 0$$

Exercise

$$\widehat{\mu} = \bar{y}_{\cdot\cdot} \qquad \text{and} \qquad \widehat{\tau}_t = \bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot} \qquad \text{and} \qquad \widehat{\beta}_b = \bar{y}_{b\cdot} - \bar{y}_{\cdot\cdot}$$

Two-Way ANOVA

Decomposition

$$\underbrace{\sum_{t=1}^{T}\sum_{b=1}^{B}(y_{bt}-\bar{y}_{\cdot\cdot})^2}_{\text{SS}_{total}} = \underbrace{\sum_{t=1}^{T}\sum_{b=1}^{B}[y_{bt}-(\bar{y}_{\cdot t}-\bar{y}_{\cdot\cdot})-(\bar{y}_{b\cdot}-\bar{y}_{\cdot\cdot})-\bar{y}_{\cdot\cdot}]^2}_{\text{SS}_{err}}$$
$$+\underbrace{\sum_{t=1}^{T}\sum_{b=1}^{B}(\bar{y}_{\cdot t}-\bar{y}_{\cdot\cdot})^2}_{\text{SS}_{treat}}+\underbrace{\sum_{t=1}^{T}\sum_{b=1}^{B}(\bar{y}_{b\cdot}-\bar{y}_{\cdot\cdot})^2}_{\text{SS}_{block}}$$

F-Test

$$\frac{\text{SS}_{treat}/(T-1)}{\text{SS}_{err}/[(T-1)(B-1)]} \overset{H_0}{\sim} F_{(T-1,(T-1)(B-1))}$$

Remark $N = TB$, $\dim(M_A) = 1+(T-1)+(B-1)$, $\dim(M_0) = 1+(B-1)$.

Likewise, Nothing Special

Start with

$$\underbrace{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2}_{\text{denominator of } F} = \underbrace{\sum_t \sum_b [y_{bt} - \overset{\widehat{\mu}}{\downarrow}\bar{y}_{\cdot\cdot} - \overbrace{(\bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot})}^{\widehat{\tau}_t} - \overbrace{(\bar{y}_{b\cdot} - \bar{y}_{\cdot\cdot})}^{\widehat{\beta}_b}]^2}_{\text{SS}_{err}}$$

and

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_0\|^2 = \sum_t \sum_b [y_{bt} - \overset{\widehat{\mu}}{\downarrow}\bar{y}_{\cdot\cdot} - \overbrace{(\bar{y}_{b\cdot} - \bar{y}_{\cdot\cdot})}^{\widehat{\beta}_b}]^2,$$

can show (exercise)

$$\underbrace{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_0\|^2 - \|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2}_{\text{numerator of } F} = \underbrace{\sum_t \sum_b (\bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot})^2}_{\text{SS}_{treat}}.$$

Special Case: $T = 2$

Exercise Show that, when $T = 2$, the two-way ANOVA F-test is the same as the matched-pair t-test:

$$\frac{\bar{d}}{s/\sqrt{B}} \overset{H_0}{\sim} t_{B-1},$$

where

$$d_b = y_{b1} - y_{b2}, \qquad s = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(d_b - \bar{d})^2}.$$

Hint Specifically, show that

$$\left(\frac{\bar{d}}{s/\sqrt{B}}\right)^2 = \frac{\mathrm{SS}_{treat}/(T-1)}{\mathrm{SS}_{err}/[(T-1)(B-1)]},$$

where now $T = 2$.