

Confounding

Stratification, Standardization, and Simpson's Paradox

Yen-Yi Ho

Department of Statistics, University of South Carolina

Section 1

Part I: What Confounding Is

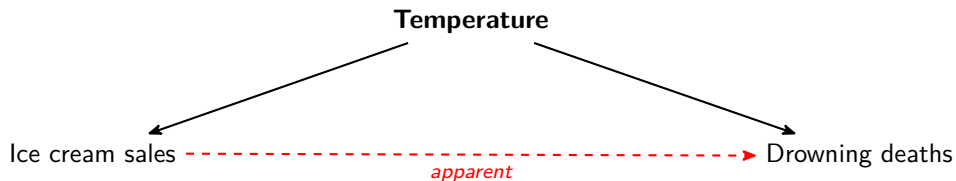
A Finding From a Public Database

A researcher looked into a public database from a southern city, and found a strong association between **ice-cream consumption** and the **number of drowning deaths** over the same period.

He concluded that ice cream could cause drowning.

Do you agree? Why do you think the researcher found such an association?

What Went Wrong



A **confounder** Z is a variable that is

- ① associated with the explanatory variable X ;
- ② **independently** associated with the response Y ;
- ③ **not** on the causal pathway from X to Y .

Hot days sell ice cream. Hot days also send people into the water.

The SUSHI Gene (I)

Once upon a time, an ethnogeneticist decided to figure out why some people eat with chopsticks and others do not.

His experiment was simple. He rounded up several hundred students from a local university, asked them how often they used chopsticks, then collected buccal DNA samples and mapped them for a series of anonymous and candidate genes.

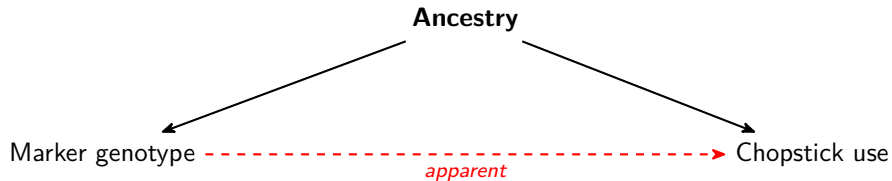
The SUSHI Gene (II)

The results were astounding. One of the markers, located right in the middle of a region previously linked to several behavioral traits, showed a huge correlation to chopstick use — enough to account for nearly half of the observed variance.

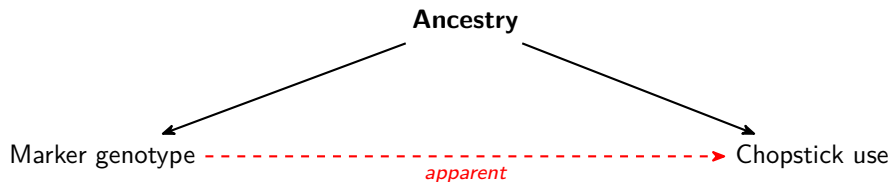
When the experiment was repeated with students from a different university, precisely the same marker lit up.

Eureka! The delighted scientist popped a bottle of champagne and submitted an article heralding the discovery of the *successful-use-of-selected-hand-instruments* gene (**SUSHI**).

The SUSHI Gene (III)



The SUSHI Gene (III)



To control for the potential confounding effect, we must **stratify by ancestry group** and perform the analysis **within** each group.

This is the central move of the whole lecture: **compare like with like**.

Section 2

Part II: Simpson's Paradox

Two Batters

	First Half	Second Half	Whole Season
Player 1	4/10 (.400)	25/100 (.250)	29/110 (.264)
Player 2	35/100 (.350)	2/10 (.200)	37/110 (.336)

Player 1 has the better average in **the first half** and in **the second half**, yet the **worse average over the season**.

Look at where the at-bats are: Player 1 took 100 of his 110 at-bats in the half where *everybody* hit badly.

Simpson's Paradox Is Not Paradoxical

$$\begin{aligned}a/b &< c/d \\ e/f &< g/h \\ (a + e)/(b + f) &> (c + g)/(d + h)\end{aligned}$$

These three statements are perfectly consistent. Nothing is broken.

Statistically: **the apparent relationship between two variables can change – even reverse – in the light or absence of a third.**

Berkeley Admissions: The Aggregate

```
data(UCBAdmissions)
apply(UCBAdmissions, c(1, 2), sum)
```

	Gender	
Admit	Male	Female
Admitted	1198	557
Rejected	1493	1278

```
tot <- apply(UCBAdmissions, c(1, 2), sum)
round(tot["Admitted", ] / colSums(tot), 3) # acceptance rate
```

	Male	Female
	0.445	0.304

Men are admitted at 44.5%, women at 30.4%. A 14-point gap.

Berkeley Admissions: By Department

```
round(apply(UCBAdmissions, 3, function(x)
  c(Male = x[1] / sum(x[1:2]), Female = x[3] / sum(x[3:4]))), 2)
```

	Dept					
Admit	A	B	C	D	E	F
Male	0.62	0.63	0.37	0.33	0.28	0.06
Female	0.82	0.68	0.34	0.35	0.24	0.07

In **four of six** departments women are admitted at the *higher* rate.

Departments A and B are easy to get into, and mostly men applied to them. Departments C–F are hard, and mostly women applied to them. Department is a confounder of the gender–admission association.

The Strategy

Variables correlated with **both** the explanatory and the response variable can distort the estimated effect.

One strategy to adjust for a confounder:

- 1 **Stratify** by the confounder;
- 2 estimate the effect **within** each stratum;
- 3 **recombine** the stratum-specific estimates — with appropriate weights.

Step 3 is the one people skip, and it is where the arithmetic lives.

Caution: unnecessary stratification reduces precision.

Section 3

Part III: Doing the Adjustment

The Question

NHANES, 5,397 adults aged 20–79. Does systolic blood pressure differ between respondents who report being physically active and those who do not?

```
inactive <- df$PhysActive == "No"
mu_i <- mean(df$BPSysAve[inactive])    # not active
mu_a <- mean(df$BPSysAve[!inactive])   # active
c(inactive = mu_i, active = mu_a, difference = mu_i - mu_a)
```

inactive	active	difference
121.365874	119.090000	2.275874

```
t.test(BPSysAve ~ PhysActive, data = df)$p.value
```

```
[1] 2.829168e-07
```

A difference of 2.28 mmHg, with $p \approx 2.8 \times 10^{-7}$. Case closed?

Step 1: Is There a Candidate Confounder?

Age is the obvious suspect. It has to clear **both** conditions.

Condition 1 — associated with the grouping:

```
tapply(df$Age, df$PhysActive, mean)
```

No	Yes
48.42845	43.79133

Condition 2 — associated with the outcome:

```
c(cor_Age_SBP = cor(df$Age, df$BPSysAve))
```

```
cor_Age_SBP  
0.3639044
```

Inactive respondents are about **4.6 years older**, and blood pressure climbs with age ($r = 0.36$). Both conditions hold.

Step 1, Seen Directly

```
round(rbind(`% inactive` = 100 * tapply(inactive, df$AgeDecade, mean),  
          `mean SBP`     = tapply(df$BPSysAve, df$AgeDecade, mean)), 1)
```

	20-29	30-39	40-49	50-59	60-69	70+
% inactive	33.8	39.7	45.2	47.1	49.1	63.6
mean SBP	113.2	115.5	118.6	123.4	126.7	131.7

As we move across the decades, the inactive share climbs from **34% to 64%** *and* mean pressure climbs from **113 to 132 mmHg**.

The “inactive” group is simply an **older** group. Some of the 2.28 mmHg could be due to age, not activity.

Step 2: Stratify

```
dec <- levels(df$AgeDecade)
tab <- data.frame(decade = dec, n = NA_integer_,
                  ybar_i = NA_real_, ybar_a = NA_real_, diff = NA_real_)
for (i in seq_along(dec)) {
  s <- df[df$AgeDecade == dec[i], ]
  a <- s$PhysActive == "Yes"
  tab$n[i] <- nrow(s)
  tab$ybar_i[i] <- mean(s$BPSysAve[!a]) # inactive mean
  tab$ybar_a[i] <- mean(s$BPSysAve[a])  # active mean
  tab$diff[i] <- tab$ybar_i[i] - tab$ybar_a[i] # within-decade difference
}
tab$w <- tab$n / sum(tab$n) # stratum weights
```

tab\$diff holds six **like-with-like** comparisons: within a decade, everyone is roughly the same age, so age cannot drive the difference.

Step 2: The Six Comparisons

Decade	n	Inactive	Active	Difference	w
20-29	1024	111.76	113.88	-2.12	0.1897
30-39	1018	114.78	115.89	-1.11	0.1886
40-49	1115	118.96	118.36	0.60	0.2066
50-59	1047	126.45	120.61	5.84	0.1940
60-69	753	127.36	126.08	1.28	0.1395
70+	440	130.18	134.22	-4.05	0.0815

The crude difference said inactive respondents sit 2.28 mmHg **higher**. Here **three of the six decades run the other way** — in those decades the active group has the higher mean.

The Key Question

The six within-decade differences average roughly half a point.

The crude difference was 2.28.

Where did the crude number come from, then?

Answer: the crude difference is **also** a weighted average of the strata — but it uses **two different sets of weights**, one per group.

Why the Crude Number Is Larger

Write $\bar{y}_s^I, \bar{y}_s^A$ for the stratum means and let π_s^I, π_s^A be each **group's own** age distribution.

$$\Delta_{\text{crude}} = \underbrace{\sum_s \pi_s^I \bar{y}_s^I}_{\text{inactive mean}} - \underbrace{\sum_s \pi_s^A \bar{y}_s^A}_{\text{active mean}}$$

```
pi_i <- tapply(inactive, df$AgeDecade, sum) / sum(inactive)
pi_a <- tapply(!inactive, df$AgeDecade, sum) / sum(!inactive)
round(100 * rbind(inactive = pi_i, active = pi_a), 1)
```

	20-29	30-39	40-49	50-59	60-69	70+
inactive	14.4	16.9	21.0	20.6	15.4	11.7
active	22.6	20.5	20.4	18.5	12.8	5.3

The inactive group carries **11.7%** of its weight in the 70+ decade; the active group carries **5.3%**. The groups are being averaged over *different populations*.

The Fix: Use One Common Set of Weights

$$\Delta_{\text{adj}} = \sum_s w_s (\bar{y}_s^I - \bar{y}_s^A) = \sum_s w_s \bar{y}_s^I - \sum_s w_s \bar{y}_s^A$$

Both groups are re-weighted to the **same** age distribution w_s — here the whole sample's, $w_s = n_s/n$. This is **direct standardization**.

The second form says the same thing a different way: take each group's mean as it *would have been* if that group had the age profile of the full sample, then subtract.

Step 3: Recombine

```
adjusted <- sum(tab$w * tab$diff)
c(crude = mu_i - mu_a, adjusted = adjusted)
```

```
      crude  adjusted
2.2758740 0.4955471
```

```
# identical, via standardized group means
```

```
c(std_inactive = sum(tab$w * tab$ybar_i),
   std_active   = sum(tab$w * tab$ybar_a),
   difference    = sum(tab$w * tab$ybar_i) - sum(tab$w * tab$ybar_a))
```

```
std_inactive  std_active  difference
120.3472415   119.8516944    0.4955471
```

What Adjustment Did

Quantity	mmHg
Crude difference (inactive — active)	+2.28
Age-adjusted difference	+0.50

Seeing It

```
long <- data.frame(
  decade = rep(tab$decade, 2),
  group   = rep(c("Not active", "Active"), each = nrow(tab)),
  mean    = c(tab$ybar_i, tab$ybar_a))

ggplot(long, aes(decade, mean, colour = group, group = group)) +
  geom_line(linewidth = 1) + geom_point(size = 2.5) +
  scale_colour_manual(values = c("Not active" = "#C44E52",
                                "Active"      = "#4C72B0")) +
  labs(title = "Mean systolic BP by age decade and activity",
       x = "Age decade", y = "Mean SBP (mmHg)", colour = NULL)
```

Confounding

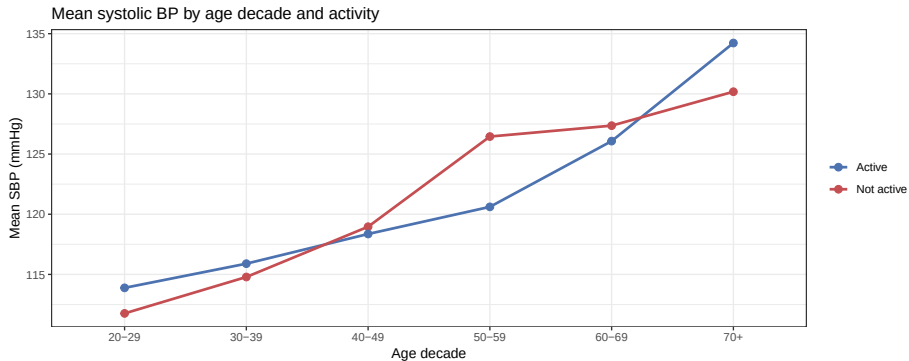


Figure 1

Both lines climb steeply with age; the gap between them is small and switches sides. **The steep climb is what the crude comparison was picking up.**

Is What Remains Real?

Permute the activity labels **within each decade**, so age is held fixed by construction, and rebuild the adjusted difference each time.

```
set.seed(42)
B <- 1000
null <- numeric(B)
for (b in 1:B) {
  lab <- ave(df$PhysActive, df$AgeDecade, FUN = sample) # shuffle within decade
  d <- numeric(length(dec))
  for (i in seq_along(dec)) {
    k <- df$AgeDecade == dec[i]; a <- lab[k] == "Yes"
    d[i] <- mean(df$BPSysAve[k][!a]) - mean(df$BPSysAve[k][a])
  }
  null[b] <- sum(tab$w * d)
}
(sum(abs(null) >= abs(adjusted)) + 1) / (B + 1)
```

The Punchline

	Difference	p -value
Crude	+2.28	2.8×10^{-7}
Age-adjusted	+0.50	≈ 0.23

A p -value of 2.8×10^{-7} .

- A p -value asks: *could chance alone produce a difference this large?*
- Confounding asks: *Could something else explain the difference?*

Choosing the Weights

$\Delta_{\text{adj}} = \sum_s w_s \Delta_s$ needs a choice of w_s .

Common choices

- $w_s = n_s/n$ — the sample itself (used here)
- An **external standard population**, e.g. the 2000 US census age distribution, so studies are comparable
- $w_s \propto 1/\widehat{\text{Var}}(\Delta_s)$ — most precise

Things to watch

- Report which standard you used — the number changes with it
- Thin strata give unstable Δ_s
- Too many strata: precision falls
- Too few: confounding survives inside a stratum

What Adjustment Does Not Do

- It adjusts for **the confounders you measured**. Gender, diet, income, and the illness that stopped someone exercising are all still in there.
- **Reverse causation** is untouched: perhaps illness caused the inactivity rather than the other way round.
- Adjusting for a variable **on the causal pathway** removes part of the very effect you want (condition 3 — do not adjust for a mediator).
- It gives an association within strata. It does **not** make the study an experiment.

Regression does the same job by a different route: $\text{lm}(\text{BPSysAve} \sim \text{PhysActive} + \text{Age})$ fits one common slope for age rather than splitting into bins. Same idea, stronger assumptions.

Section 4

Part IV: Where Evidence Comes From

A Cure for Deafness

On 15 July 1911, 65-year-old Mrs. Jane Decker was struck by lightning while in her house. She had been deaf since birth, but after being struck, she recovered her hearing.

Is this compelling evidence that lightning cures deafness?

Hierarchy for Scientific Evidence

- **Observational study**

- Anecdotal evidence: a short story, or an example of an interesting event
- Retrospective case-control studies
- Prospective cohort studies

- **Experiment**

- Randomized clinical trials

Confounding is a threat at every observational level. Randomization is the only design that handles the confounders you **did not think to measure**.

Smoking and Lung Cancer

Smoking → lung cancer

- **Anecdotal:** my uncle smoked and he got lung cancer.
- **Retrospective case-control:** 100 lung cancer patients and 100 normal controls were enrolled; then smoking history for each participant was examined.
- **Prospective cohort:** 100 smokers and 100 nonsmokers were enrolled, then followed up to determine whether they developed lung cancer.
- **Randomized trial:** participants were randomized to receive low-nicotine or regular cigarettes, then followed up to determine whether they developed lung cancer.

Randomization makes the groups alike **in every variable at once** — measured or not — which is exactly what stratification cannot promise.

Key Takeaways

Recognizing it

- A confounder is associated with X , independently with Y , and is not a mediator
- Simpson's paradox is the extreme case: the sign flips
- Always ask what the two groups differ in *besides* the label

Doing something about it

- Stratify, estimate within strata, recombine
- $\Delta_{\text{adj}} = \sum_s w_s \Delta_s$, one common weight vector
- Report the standard population you used
- A small p -value is no defense

Compare like with like — then say which "like" you chose.

References and Further Reading

- Rothman, K.J., Greenland, S., Lash, T.L. *Modern Epidemiology*, 3rd ed. — confounding and standardization
- Irizarry, R.A. *Introduction to Data Science*
 - Association is not causation: <https://rafalab.dfci.harvard.edu/dsbook-part-1/>
- Bickel, P.J., Hammel, E.A., O'Connell, J.W. (1975) "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187:398–404
- Hanis, C.L. et al. — the chopstick-gene cautionary tale, as retold by Hodge, S.E. (1994)
- ?UCBAdmissions, ?NHANES in R