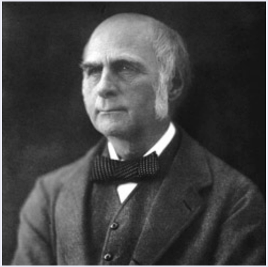
Chapter 1

Introduction

Sir Francis Galton (1822-1911)

- Galton was a polymath who made important contributions in many fields of science, including meteorology (the anti-cyclone and the first popular weather maps), statistics (regression and correlation), psychology (synesthesia), biology (the nature and mechanism of heredity), and criminology (fingerprints)
- He first introduced the use of questionnaires and surveys for collecting data on human communities.



Karl Pearson (1857 - 1936)

- student of Francis Galton
- He has been credited with establishing the discipline of mathematical statistics, and contributed significantly to the field of biometrics, meteorology, theories of social Darwinism and eugenics
- Founding chair of department of Applied Statistics in University of London (1911), the first stat department in the world!
- Founding editor of *Biometrika*

Incomplete Data

- Due to no direct measurement
- Due to refusal / Don't know / not available
- Due to uncertainty in the measurement
- Due to design
- Due to self-selection

Example 1: No direct measurement

- A study of managers of Iowa farmer cooperatives ($n = 98$)
- Five variables
  - x_1: Knowledge (knowledge of the economic phase of management directed toward profit-making in a business and product knowledge)
  - x_2: Value Orientation (tendency to rationally evaluate means to an economic end)
  - x_3: Role Satisfaction (gratification obtained by the manager from performing the managerial role)
  - x_4: Past Training (amount of formal education)
  - y: Role performance
- We are interested in estimating parameters in the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Example 1 (Cont'd)

Measure	No. of Items	Mean	Reliability
x_1 Knowledge	26	1.38	0.6096
x_2 Value orientation	30	2.88	0.6386
x_3 Role satisfaction	11	2.46	0.8002
x_4 Past training	1	2.12	1.0000
y Role performance	24	0.0589	0.8230

Example 1 (Cont'd)

- Ordinary least squares method

$$\begin{array}{rcl}\hat{Y} & = & -0.9740 + 0.2300X_1 + 0.1199X_2 + 0.0560X_3 + 0.1099X_4 \\ & & \qquad\qquad (0.0535) \quad (0.0356) \quad (0.0375) \quad (0.0392)\end{array}$$

- Errors-in-variable estimates

$$\begin{array}{rcl}\hat{Y} & = & -1.1828 + 0.3579X_1 + 0.1549X_2 + 0.0613X_3 + 0.0715X_4 \\ & & \qquad\qquad (0.1288) \quad (0.0794) \quad (0.0510) \quad (0.0447)\end{array}$$

Reference:

Warren, White, and Fuller (1974). "An Errors-In-Variables Analysis of Managerial Role Performance", JASA, 69, p 886-893.

Example 2. Asthma Study Data (Pigott, 2001)

Variable descriptions

Variable	Definition	Possible values	Mean	N
Asthma belief	Level of confidence	1= little confidence 5= lots of confidence	4.057	154
Group	Treatment or control	0 = treatment 1 = control	0.558	154
Symsev	Severity of asthma symptoms in 2 weeks	0 = no symptoms 3 = severe symptoms	0.235	141
Reading	Standardized state reading test scores	Grade equivalent scores, from 1.10 to 8.10	3.443	79
Age		Ranging from 8 to 14	10.586	152
Gender		0 =Male 1 = Female	0.442	154
Allergy	No. of allergies	Range from 0 to 7	2.783	83

Missing Data Patterns

Symsev	Reading	Age	Allergy	# of cases	% of cases
O	O	O	O	19	12.3
M	O	O	O	1	0.6
O	M	O	O	54	35.1
O	O	O	M	56	36.4
M	M	O	O	9	5.8
M	O	O	M	1	0.6
O	M	O	M	10	6.5
O	O	M	M	2	1.3
M	M	O	M	2	1.3
				154	100.0

Example 2 (Cont'd)

Results (CC: Complete Case, ML: Maximum Likelihood)

Variable	CC analysis		ML analysis	
	B	SE	B	SE
Intercept	4.617	0.838	4.083	0.362
Trt group	-0.550	0.276	-0.132	0.112
Symsev	-0.315	0.161	-0.480	0.144
Reading	0.409	0.096	0.218	0.039
Age	-0.211	0.115	-0.089	0.043
Gender	0.198	0.189	0.084	0.104
Allergy	-0.005	0.057	0.063	0.029

Reference:

Pigott (2001). "A Review of Methods for Missing Data", *Educational Research and Evaluation*, 7, 353-383.

Example 3: 2009 Local Area Labor Force survey in Korea.

- Large scale survey with about $n = 157K$ sample households.
- Obtain the employment status: Employed, Unemployed, Not in labor force.
- To obtain response, interviewers visit the sample households up to four times. That is, the current rule allows for three follow-ups.

Example 3 (Cont'd)

Realized Responses from the Korean LF survey data

status	t=1	t=2	t=3	t=4	No response
Employment	81,685	46,926	28,124	15,992	
Unemployment	1,509	948	597	352	32,350
Not in LF	57,882	32,308	19,086	10,790	

Example 3 (Cont'd)

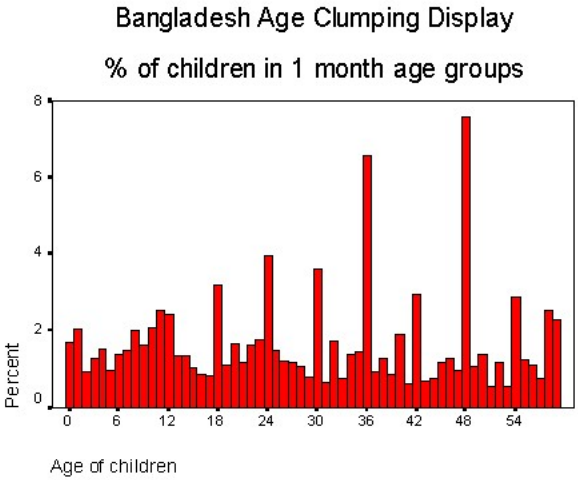
	First Response at t-th visit				No Response
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	
Response Rate (%)	42.94	24.40	14.55	8.26	9.85
Ave. Unemp. Rate (%)	1.81	1.98	2.08	2.15	?

Response propensity seems to be correlated with the unemployment rate.

Reference:

Kim, J.K. and Im, J. (2014). "Propensity score weighting adjustment with several follow-ups", *Biometrika* **101**, 439-448.

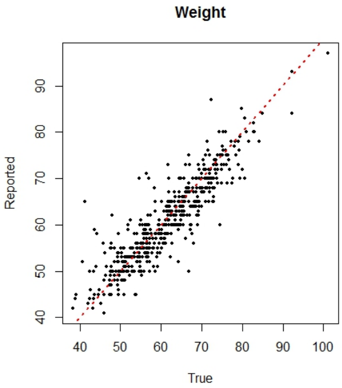
Measurement error: Age Heaping example



Measurement error: BMI data example

- Korean Longitudinal Study of Aging (KLoSA) data (http://www.kli.re.kr/klosa/en/about/introduce.jsp)
- Original sample measures height and weight from survey questions (N=9,842)
- A validation sample (n=505) is randomly selected from the original sample to obtain physical measurement for the height and weight.

Measurement error: BMI data example (Cont'd)



Planned missingness: NRI example

National Resources Inventory
(http://www.nrcs.usda.gov/wps/portal/nrcs/main/national/technical/nra/nri/)

1997	2000	2001	2002	2003	2004
✓	✓	✓	✓	✓	✓
✓		✓			
✓			✓		
✓				✓	
✓					✓
✓					
✓					
✓					

Planned missingness: Split questionaire design

Pattern	x	y_1	y_2	y_3	Cost	Sample Size
1	✓	✓			c_1	n_1
2	✓		✓		c_2	n_2
3	✓			✓	c_3	n_3
4	✓	✓	✓		c_4	n_4
5	✓		✓	✓	c_5	n_5
6	✓	✓		✓	c_6	n_6
7	✓	✓	✓	✓	c_7	n_7

Reference:

Chipperfield and Steel (2009). "Design and Estimation for Split Questionnaire Surveys", *Journal of Official Statistics* **25**, 227–244.

Will generally use this notation throughout

$$
\begin{aligned}
Y &= \text{outcome or dependent variable} \\
X &= \text{covariate or vector of covariates} \\
R &= \text{response indicator for } Y \\
&= 1 \text{ if } Y \text{ observed, } 0 \text{ if missing}
\end{aligned}
$$

Simulating data in R

Simulate observations from a normal distribution

```
## 5 observations from N(0,1)
> rnorm(n=5, mean=0, sd=1)
[1] -0.27961336  0.88267457  0.01061641 -0.08252131  0.61003977

> z = rnorm(n=5, mean=0, sd=1)
> z
[1]  0.6741197 -0.3814703  1.4246447  0.2252487 -0.1592414
> zbar = mean(z)
> zbar
[1] 0.3566603

## 30 observations from N(3,5^2)
> y = rnorm(n=30, mean=3, sd=5)
```

Simulating data in R

Summarize results of 100 simulations

```
### Simulate 5 observations 100 times
> results        = matrix(0, nrow=100, ncol=2)
> colnames(results) = c("Mean", "SD")

> for (i in 1:100)
 { z    = rnorm(n=5, mean=0, sd=1)
  results[i,1] = mean(z)
  results[i,2] = sd(z) }

### Print results
> results[1:5,]
            Mean        SD
[1,] -0.08047987 0.8044978
[2,]  0.42806792 0.4017826
[3,]  0.86330499 1.7292280
[4,] -0.53925212 1.1389417
[5,] -0.07935075 0.6154337
```

Simulating data in R

```
> results[1:5,]
            Mean        SD
[1,] -0.08047987 0.8044978
[2,]  0.42806792 0.4017826
[3,]  0.86330499 1.7292280
[4,] -0.53925212 1.1389417
[5,] -0.07935075 0.6154337

### calculate mean of individual sample means and SD's
> apply(results, 2, mean)
      Mean         SD
0.03208639 0.95688116

### standard deviation of individual sample means and SD's
> apply(results, 2, sd)
     Mean        SD
0.4985703 0.3696412
```

Simulating binary data in R

Use command `rbinom`

```
### Simulate 10 binary observations having P(R=1) = .30

> R  = rbinom(n=10, size=1, prob=.30)
> R
 [1] 0 1 0 1 0 0 1 0 1 1
> mean(R)
[1] 0.5


> R  = rbinom(n=10, size=1, prob=.30)
> R
 [1] 0 1 0 1 0 0 0 0 1 0
> mean(R)
[1] 0.3
```

Simulating incomplete data in R

1. Generate the 'full data' – in this case a sample of continuous outcomes Y
2. Generate the response indicators R – the *missing data mechanism*
   - Have to determine $P(R = 1)$
   - Can allow $P(R = 1)$ to depend on Y

Simulating incomplete data in R

Example 1. Random deletion, or missing (completely) at random.

$$
\begin{aligned}
Y &\sim N(0,1) \\
R &\sim \text{Ber}(0.5)
\end{aligned}
$$

Example 2. Deletion depends on Y such that lower values of Y are more likely to be observed. This is missing *not* at random.

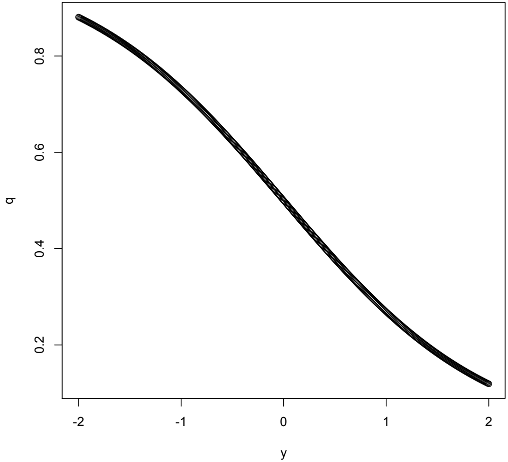
$$
\begin{aligned}
Y &\sim N(0,1) \\
R &\sim \text{Ber}\{q(Y)\}
\end{aligned}
$$

where the function $q(Y)$ is given by

$$
q(Y) = \frac{1}{1+\exp(Y)}
$$

Simulating incomplete data in R

Probability of response as a function of Y



A more general form of missing data mechanism

Can introduce a parameter that governs degree of dependence on Y

$$q(\alpha Y) = \frac{1}{1+\exp(\alpha Y)}$$

- When $\alpha = 0$, response probability does not depend on Y.
- For $\alpha \neq 0$, response probability depends on Y
- Magnitude of α governs degree of dependence

Different missing data mechanisms

The full-data model here is

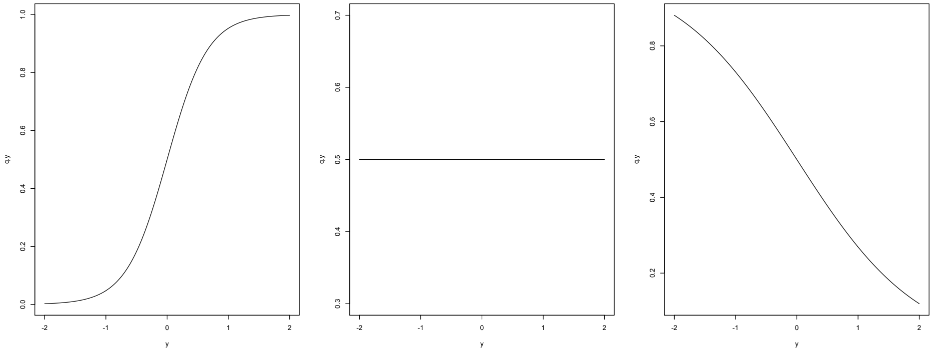
$$
\begin{aligned}
Y &\sim N(0,1) \\
R &\sim \text{Ber}\{q(\alpha Y)\}
\end{aligned}
$$

where

$$
q(\alpha Y) = \frac{1}{1+\exp(\alpha Y)}
$$

Different missing data mechanisms

These plots represent $\alpha = -3$, $\alpha = 0$, $\alpha = 1$



R Code for simulation

```
## Example 2:  nonrandom deletion
Y  = rnorm(n = 100, mean=0, sd=1)

q.Y = 1 / ( 1 + exp(Y) )
R = rbinom(n = 100, size=1, prob=q.Y)

Fulldata = cbind(Y,R)
Y.obs = Fulldata[R==1,1]

Y.obs
mean(Y)
mean(Y.obs)
mean(R)
```

R Code for simulation

```
## Simulate the process in example #2 1000 times
results = matrix(0, nrow=1000, ncol=3)
summary = matrix(0, nrow=1, ncol=3)
labels = c("mean of Y", "mean of Y.obs", "mean of R")

colnames(results) = labels
colnames(summary) = labels


# alpha controls whether R depends on Y
alpha  = 1
```

R Code for simulation

```
for (i in 1:1000)
{
Y  = rnorm(n = 100, mean=0, sd=1)
q.Y = 1 / ( 1 + exp( alpha*Y ) )
R = rbinom(n = 100, size=1, prob=q.Y)
Fulldata = cbind(Y,R)
Y.obs = Fulldata[R==1,1]
results[i,] = c( mean(Y), mean(Y.obs), mean(R) )
}

summary = apply(results, 2, mean)
summary
```

Result

```
ALPHA = -3
> summary
    mean of Y mean of Y.obs     mean of R
 0.0005652042  0.6911873446  0.4994900000

ALPHA = 0
> summary
    mean of Y mean of Y.obs     mean of R
 -0.001543965   0.001200788   0.501350000

ALPHA = 1
> summary
    mean of Y mean of Y.obs     mean of R
-0.0004493881 -0.4136889588  0.4999100000
```