Hypothesis testing with many variables

What's the problem?

Rhine paradox

Joseph Rhine – a 50's parapsychologist who hypothesized that some people have Extra-Sensory Perception

To find them, he run subjects through an experiment where they needed to guess colours of 10 hidden cards – red or blue

Rhine paradox

Joseph Rhine – a 50's parapsychologist who hypothesized that some people have Extra-Sensory Perception

To find them, he run subjects through an experiment where they needed to guess colours of 10 hidden cards – red or blue

H_0: the subject is guessing colours at random
H_1: the subject has insights about cards' colours
Statistics – number of correctly guessed card colours:

$$P(t \geq 9|H_0) = 11 \times 0.5^{10} \approx 0.01$$

$t = 9$ gets $p \approx 0.01$ – H_0 could be rejected.

Rhine paradox

1000 subjects went through the selection procedure. Nine guessed 9 of 10 colours, two – all ten.

None of them was confirmed to have ESP in the subsequent experiments.

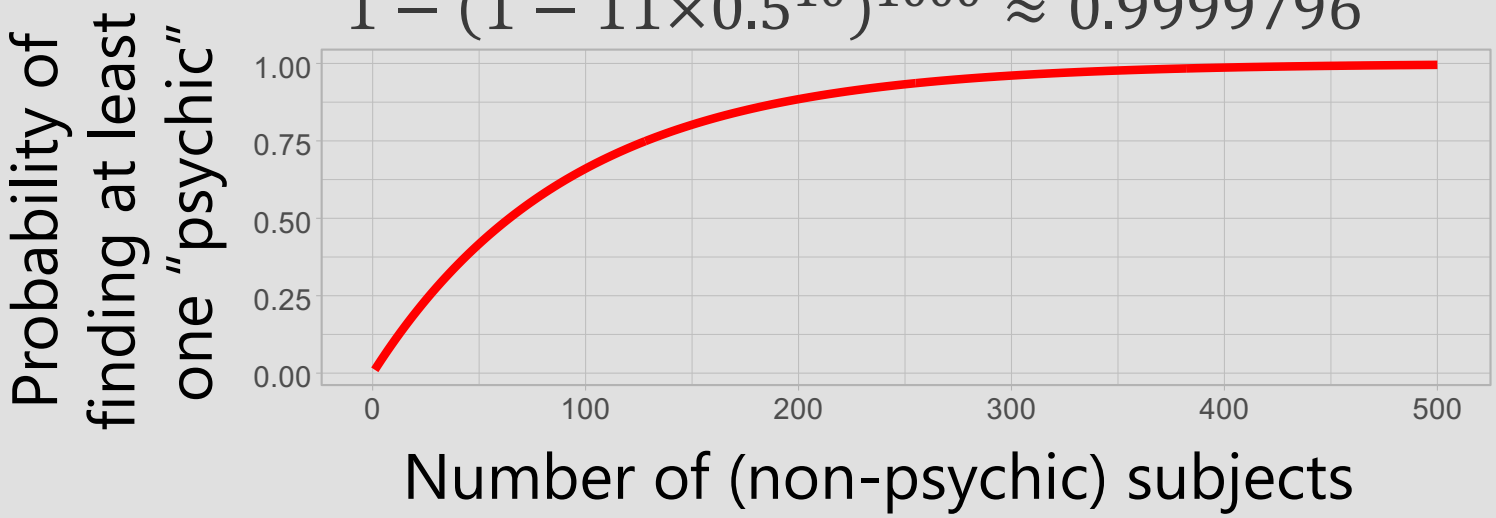
Rhine paradox

1000 subjects went through the selection procedure. Nine guessed 9 of 10 colours, two – all ten.

None of them was confirmed to have ESP in the subsequent experiments.

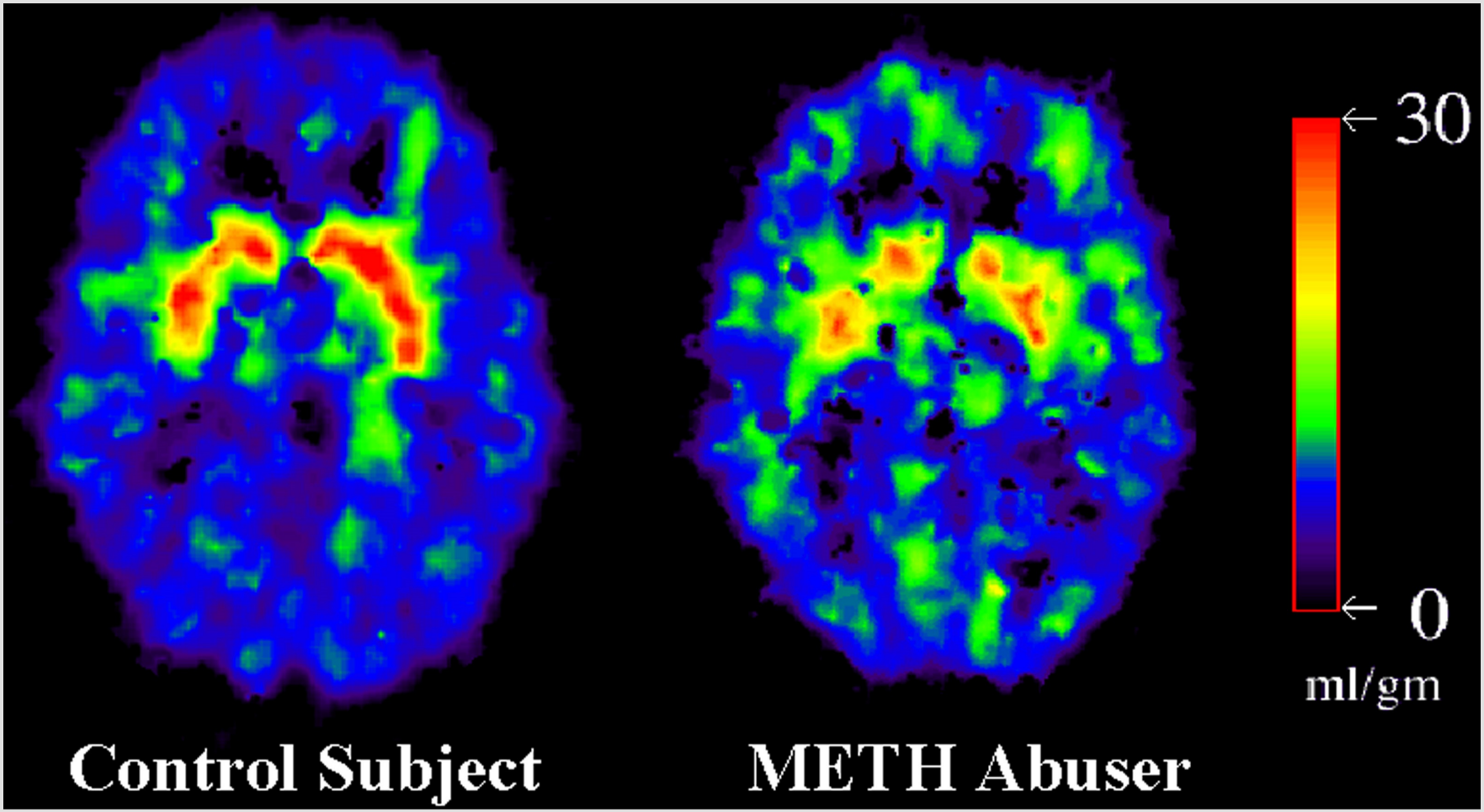
Probability that at least one of 1000 would randomly guess at least 9 of 10 colours:

$$1-(1-11\times 0.5^{10})^{1000}\approx 0.9999796$$

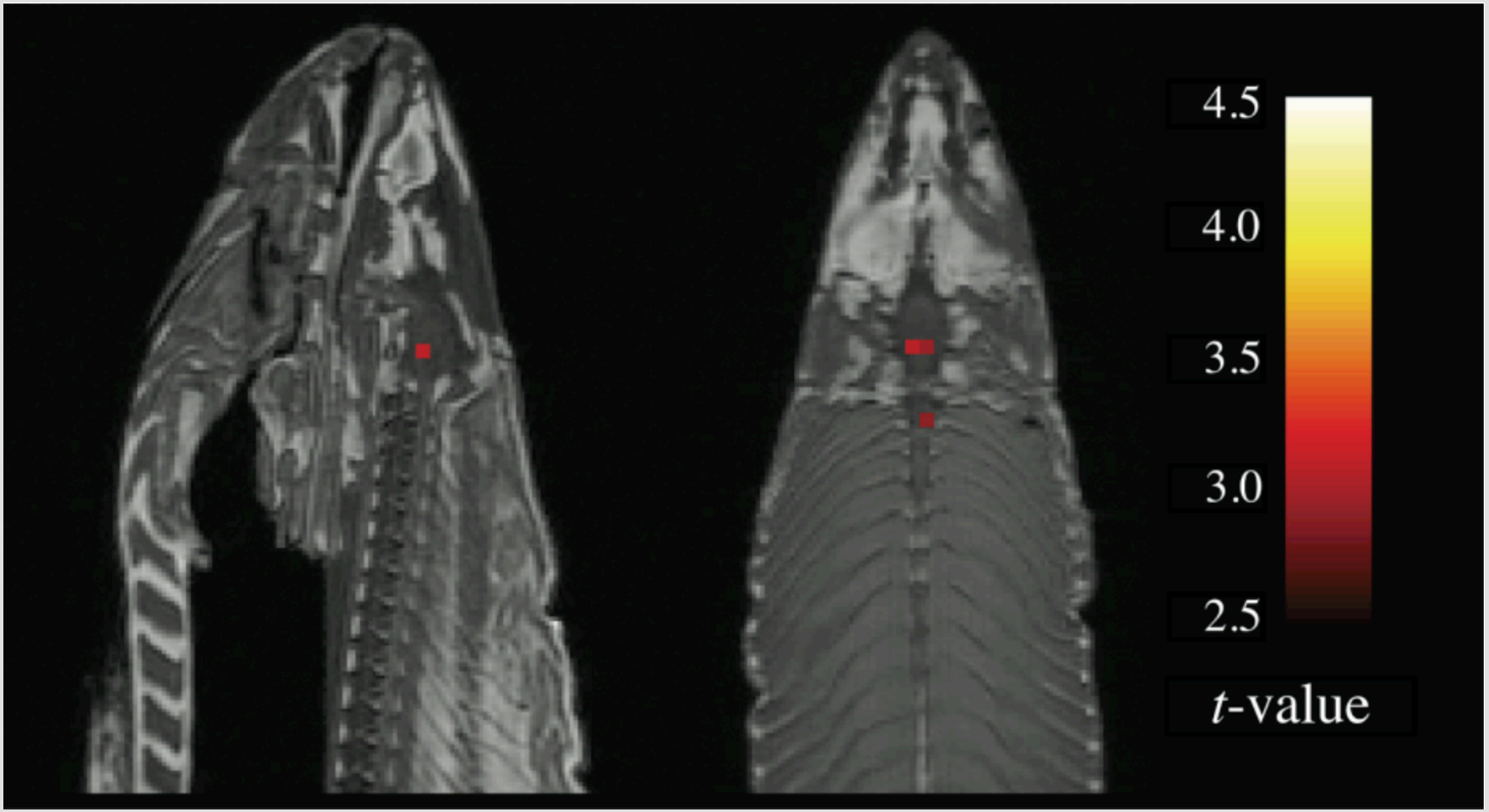


Rajaraman, Ullman, cs345a, 2009

Neurosciences



Neurosciences



Bennet et al, 2009

Testing single hypothesis

sample:	$X^n = (X_1, \dots, X_n), \mathrm{X} \sim F_X \in \Omega$
null hypothesis:	$H_0: F_X \in \omega, \omega \in \Omega$
alternative hypothesis:	$H_1: F_X \notin \omega$
statistic:	$T(X^n) \sim F(x)$ when $F_X \in \omega$ $\nsim F(x)$ when $F_X \notin \omega$
observed sample:	$x^n = (x_1, \dots, x_n)$
observed statistic:	$t = T(x^n)$
p-value:	$p(t)$ – probability of getting $T = t$ or more extreme under H_0

Null hypothesis is rejected when $p \leq \alpha$,
α – significance level.

Type I and II errors

	H_0 true	H_0 false
H_0 accepted	👍	Type II error
H_0 rejected	Type I error	👍

Tests are constructed to bound the probability of type I error by significance level α.

Testing multiple hypotheses

samples: $\mathbb{X} = \left(X_1^{n_1} \ldots, X_m^{n_m}\right), X_i \sim F_i \in \Omega_i$

null hypotheses: $H_i: F_i \in \omega_i, \omega_i \in \Omega_i$

alternative hypotheses: $H_i': F_i \notin \omega_i$

statistics: $T_i\left(X_i^{n_i}\right)$

observed statistics: $t_i = T_i\left(x_i^{n_i}\right)$

p-values: $p_i, i = 1, \ldots, m$

Testing multiple hypotheses

$\boldsymbol{M}_0 = \{i: H_i \text{ is true}\}$, $|\boldsymbol{M}_0| = m_0$
$\boldsymbol{R} = \{i: H_i \text{ is rejected}\}$, $|\boldsymbol{R}| = R$;
$V = |\boldsymbol{M}_0 \cap \boldsymbol{R}|$ – number of type I errors

	True H_is	False H_is	Total
Accepted H_is	U	T	$M - R$
Rejected H_is	V	S	R
Total	m_0	$\mathrm{m} - \mathrm{m}_0$	m

We have to do something about V.

Familywise error rate

FWER

	True H_is	False H_is	Total
Accepted H_is	U	T	$M - R$
Rejected H_is	V	S	R
Total	m_0	$\mathrm{m} - \mathrm{m}_0$	m

Familywise (type I) **error rate**:

$$\mathrm{FWER} = P(V > 0)$$

Controlling FWER at level α:

$$\mathrm{FWER} = P(V > 0) \leq \alpha$$

How?

FWER

Familywise (type I) **error rate**:

$$\mathrm{FWER} = P(V > 0)$$

Controlling FWER at level α:

$$\mathrm{FWER} = P(V > 0) \leq \alpha$$

How?

$\alpha_1, \ldots, \alpha_m$ – significance levels for $H_1, \ldots, H_m$
We have to select them to ensure $\mathrm{FWER} \leq \alpha$.

Bonferroni correction

Bonferroni method:

$$\alpha_1 = \cdots = \alpha_m = \frac{\alpha}{m}$$

Comparing α_i and p_i is the same as comparing original α and **adjusted p-value**

$$\tilde{p}_i = \min(1, mp_i)$$

H_i is rejected when $\tilde{p}_i \leq \alpha$.

Bonferroni correction

Theorem. If H_i is rejected when $p_i \leq \alpha/m$, then $\mathrm{FWER} \leq \alpha$.

Proof.

$$\mathrm{FWER} = P(V > 0) = P\left(\bigcup_{i \in \boldsymbol{M}_0} \left\{p_i \leq \frac{\alpha}{m}\right\}\right) \leq$$

$$\leq \sum_{i \in \boldsymbol{M}_0} P\left(p_i \leq \frac{\alpha}{m}\right) \leq$$

$$\leq \sum_{i \in \boldsymbol{M}_0} \frac{\alpha}{m} = \frac{m_0}{m}\alpha \leq \alpha$$

Model experiment

50 samples from $N(1,1)$, 150 samples from $N(0,1)$, $n = 20$

$H_i: \mathbb{E}X_i = 0, H_i': \mathbb{E}X_i \neq 0$, one sample t-test

Model experiment

50 samples from $N(1,1)$, 150 samples from $N(0,1)$, $n = 20$

$H_i: \mathbb{E}X_i = 0, H_i': \mathbb{E}X_i \neq 0$, one sample t-test

No corrections:

	True H_is	False H_is	Total
Accepted H_is	143	0	143
Rejected H_is	7	50	57
Total	150	50	200

Model experiment

50 samples from $N(1,1)$, 150 samples from $N(0,1)$, $n = 20$
$H_i: \mathbb{E}X_i = 0, H_i': \mathbb{E}X_i \neq 0$, one sample t-test

No corrections:

	True H_is	False H_is	Total
Accepted H_is	143	0	143
Rejected H_is	7	50	57
Total	150	50	200

Bonferroni correction:

	True H_is	False H_is	Total
Accepted H_is	150	19	169
Rejected H_is	0	31	31
Total	150	50	200

Can we do better?

Bonferroni method:

$$\alpha_1 = \cdots = \alpha_m = \frac{\alpha}{m}$$

A more powerful method is possible if we allow α_is to vary.

Step-down methods

Sorted p-values:

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

$H_{(1)}, \ldots, H_{(m)}$ – corresponding hypotheses

Step-down methods

Sorted p-values:

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

$H_{(1)}, \ldots, H_{(m)}$ – corresponding hypotheses

Step-down procedure:

1. If $p_{(1)} > \alpha_1$, accept $H_{(1)}, \ldots, H_{(m)}$ and stop; otherwise reject $H_{(1)}$ and continue
2. If $p_{(2)} > \alpha_2$, accept $H_{(2)}, \ldots, H_{(m)}$ and stop; otherwise reject $H_{(2)}$ and continue
3. …

Holm's method

Holm's method – a step-down procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha$$

Holm's method

Holm's method – a step-down procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha$$

Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \max\left((m-1+1)p_{(i)}, \tilde{p}_{(i-1)}\right)\right)$$

Holm's method

Holm's method – a step-down procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \ldots, \alpha_i = \frac{\alpha}{m-i+1}, \ldots, \alpha_m = \alpha$$

Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \max\left((m-1+1)p_{(i)}, \tilde{p}_{(i-1)}\right)\right)$$

Holm's method

Holm's method – a step-down procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \ldots, \alpha_i = \frac{\alpha}{m-i+1}, \ldots, \alpha_m = \alpha$$

Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \max\left((m-1+1)p_{(i)}, \tilde{p}_{(i-1)}\right)\right)$$

Holm's method

Holm's method – a step-down procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha$$

Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \max\left((m-1+1)p_{(i)}, \tilde{p}_{(i-1)}\right)\right)$$

Holm's method

Holm's method – a step-down procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha$$

Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \max\left((m-1+1)p_{(i)}, \tilde{p}_{(i-1)}\right)\right)$$

- $\mathrm{FWER} \leq \alpha$ is guaranteed

Holm's method

Holm's method – a step-down procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha$$

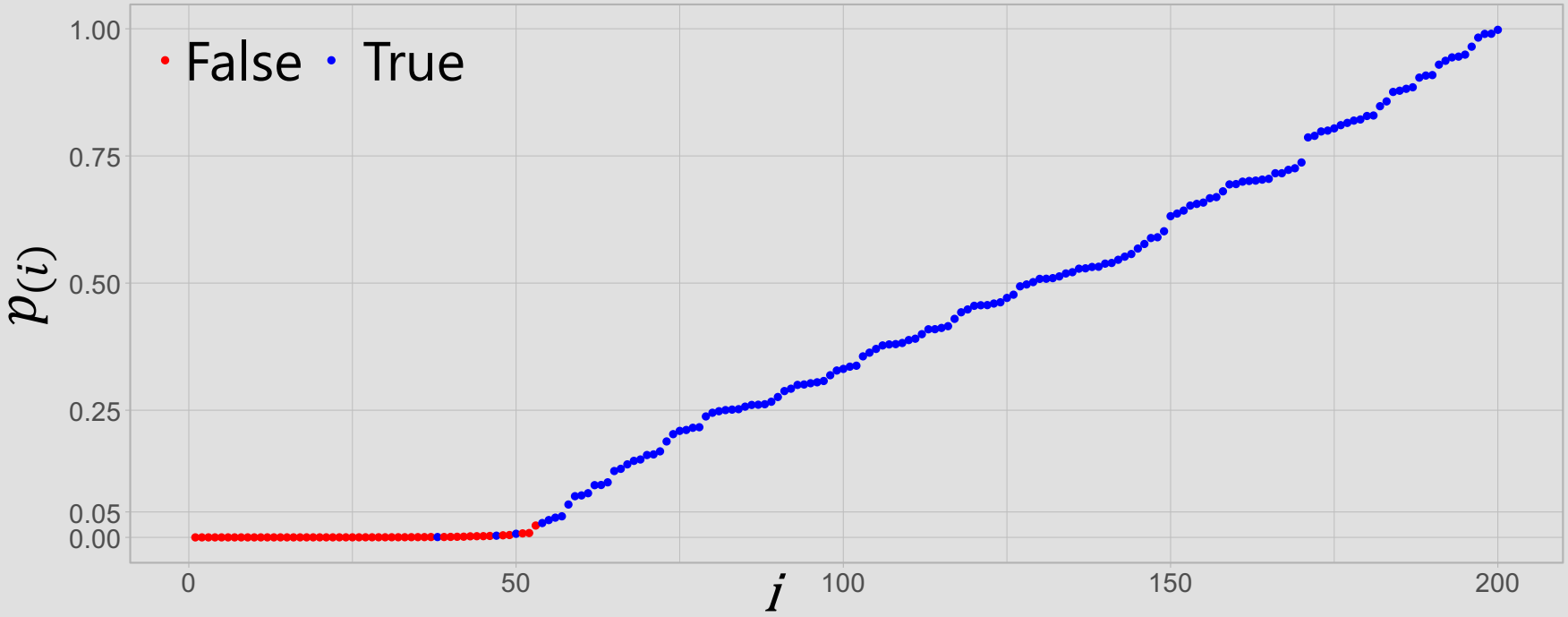
Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \max\left((m-1+1)p_{(i)}, \tilde{p}_{(i-1)}\right)\right)$$

- $\mathrm{FWER} \leq \alpha$ is guaranteed
- We always reject at least as much as Bonferroni, because α_is are always not smaller

Model experiment

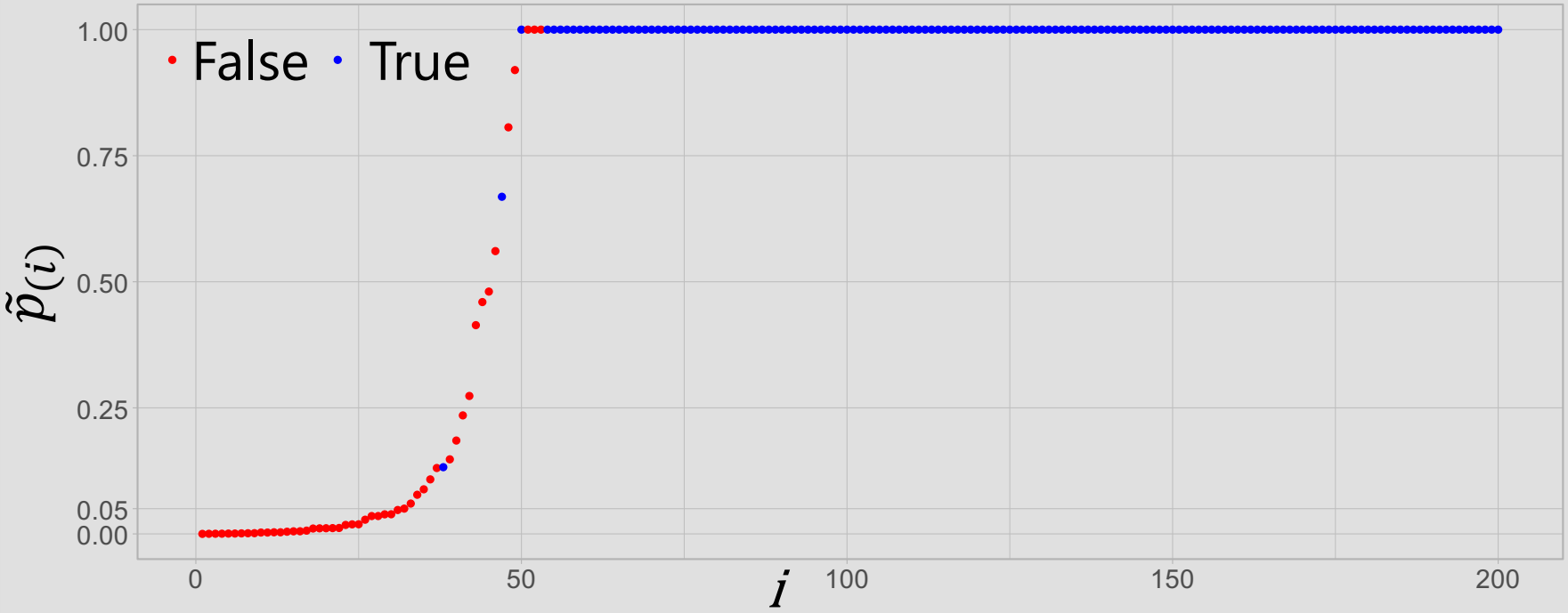
No corrections:



	True H_is	False H_is	Total
Accepted H_is	143	0	143
Rejected H_is	7	50	57
Total	150	50	200

Model experiment

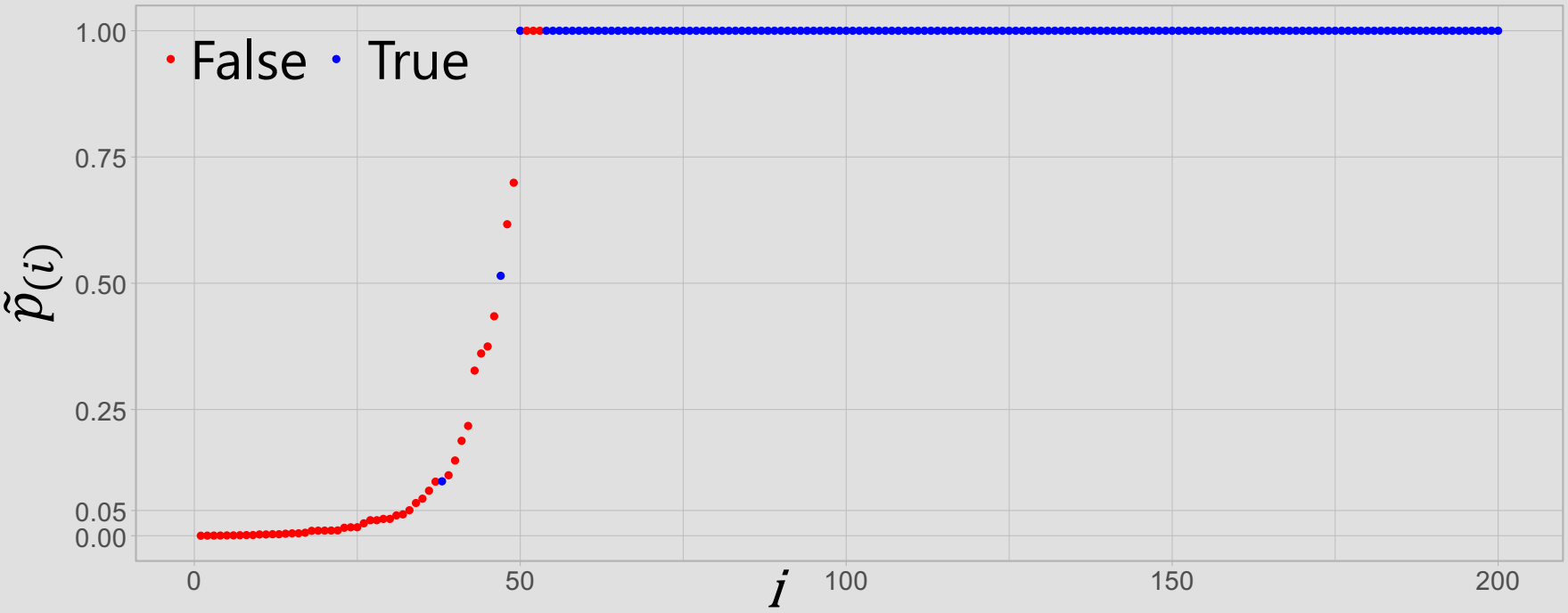
Bonferroni correction:



	True H_is	False H_is	Total
Accepted H_is	150	19	169
Rejected H_is	0	31	31
Total	150	50	200

Model experiment

Holm's method:



	True H_is	False H_is	Total
Accepted H_is	150	18	168
Rejected H_is	0	32	32
Total	150	50	200

Model experiment

Bonferroni correction:

	True H_is	False H_is	Total
Accepted H_is	150	19	169
Rejected H_is	0	31	31
Total	150	50	200

Holm's method:

	True H_is	False H_is	Total
Accepted H_is	150	18	168
Rejected H_is	0	32	32
Total	150	50	200

Takeaways about FWER

- Control FWER if it's very important not to make ANY type I error
- Use Holm's method instead of Bonferroni to reject more hypotheses FOR FREE

False discovery rate

FDR

	True H_is	False H_is	Total
Accepted H_is	U	T	$M - R$
Rejected H_is	V	S	R
Total	m_0	$\mathrm{m} - \mathrm{m}_0$	m

Familywise error rate:

$$\mathrm{FWER} = P(V > 0)$$

False discovery rate:

$$\mathrm{FDR} = \mathbb{E}\left(\frac{V}{\max(R, 1)}\right)$$

For any procedure $\mathrm{FDR} \leq \mathrm{FWER}$

Step-up methods

Sorted p-values:

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

$H_{(1)}, \ldots, H_{(m)}$ – corresponding hypotheses

Step-up methods

Sorted p-values:

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

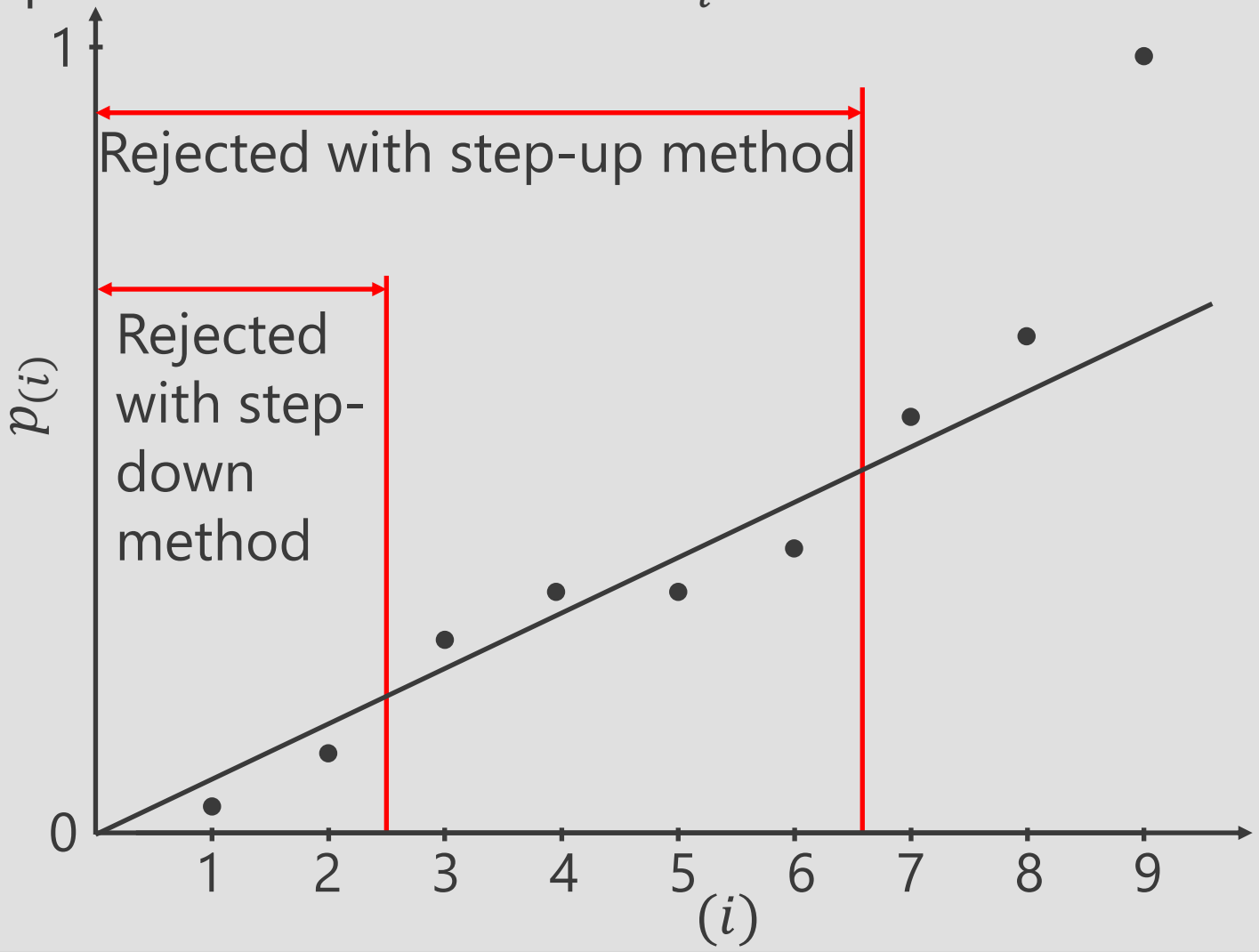
$H_{(1)}, \ldots, H_{(m)}$ – corresponding hypotheses

Step-up procedure:

1. If $p_{(m)} \leq \alpha_m$, reject $H_{(1)}, \ldots, H_{(m)}$ and stop; otherwise accept $H_{(m)}$ and continue
2. If $p_{(m-1)} \leq \alpha_{m-1}$, reject $H_{(1)}, \ldots, H_{(m-1)}$ and stop; otherwise accept $H_{(m-1)}$ and continue
3. …

Step-up methods

Step-up procedure always rejects at least as much as step-down procedure with the same α_is:



Benjamini-Hochberg's method

Benjamini-Hochberg's method – a step-up procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{2\alpha}{m}, \dots, \alpha_i = \frac{i\alpha}{m}, \dots, \alpha_m = \alpha$$

Benjamini-Hochberg's method

Benjamini-Hochberg's method – a step-up procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{2\alpha}{m}, \dots, \alpha_i = \frac{i\alpha}{m}, \dots, \alpha_m = \alpha$$

Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \frac{mp_{(i)}}{i}, \tilde{p}_{(i+1)}\right)$$

Benjamini-Hochberg's method

Benjamini-Hochberg's method – a step-up procedure with

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{2\alpha}{m}, \dots, \alpha_i = \frac{i\alpha}{m}, \dots, \alpha_m = \alpha$$

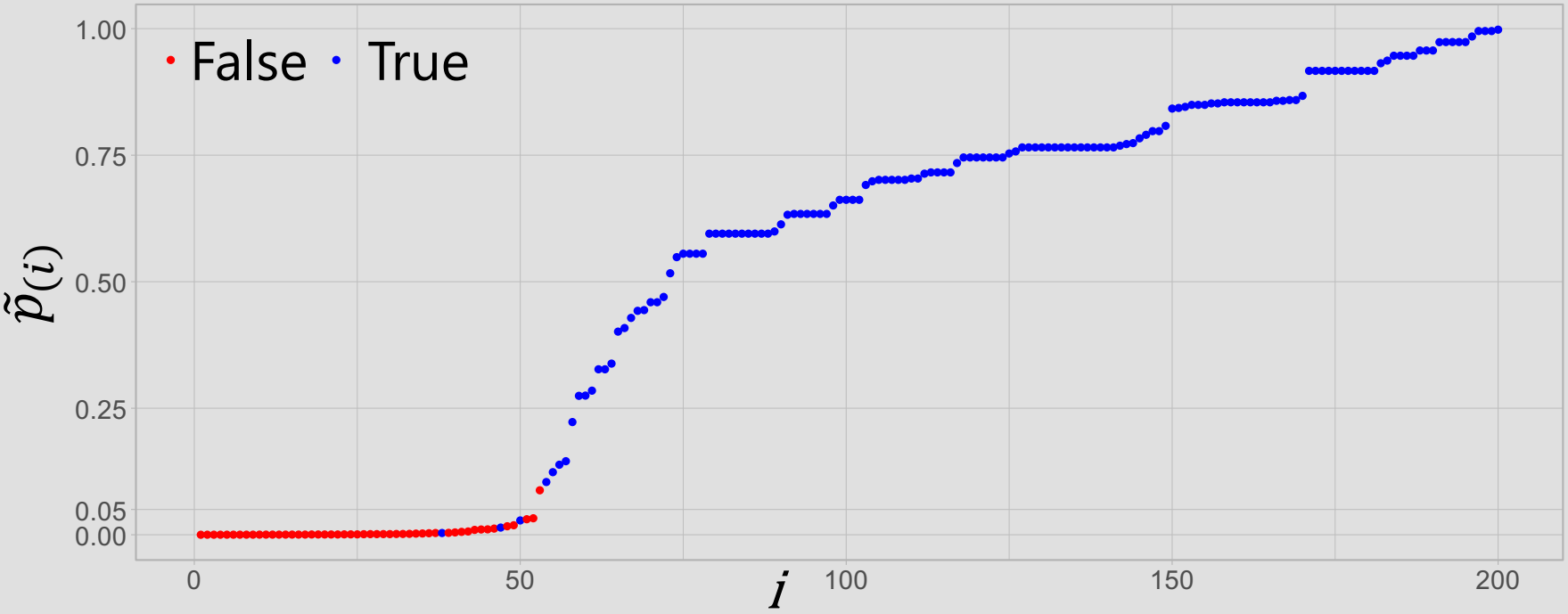
Adjusted p-values:

$$\tilde{p}_{(i)} = \min\left(1, \frac{m p_{(i)}}{i}, \tilde{p}_{(i+1)}\right)$$

- $\mathrm{FDR} \leq \alpha$ is guaranteed if $T_1, \dots, T_m$ are independent

Model experiment

Benjamini-Hochberg's method:



	True H_is	False H_is	Total
Accepted H_is	147	1	148
Rejected H_is	3	49	52
Total	150	50	200

Model experiment

Holm's method:

	True H_is	False H_is	Total
Accepted H_is	150	18	168
Rejected H_is	0	32	32
Total	150	50	200

Benjamini-Hochberg's method:

	True H_is	False H_is	Total
Accepted H_is	147	1	148
Rejected H_is	3	49	52
Total	150	50	200

Takeaways about FDR

- By controlling FDR instead of FWER, you allow some type I errors to have less type II errors
- Benjamini-Hochberg's method is proven to work only with additional assumptions about dependence between statistics
- Nevertheless, it is very widely used

Subgroup analysis

CHD treatments

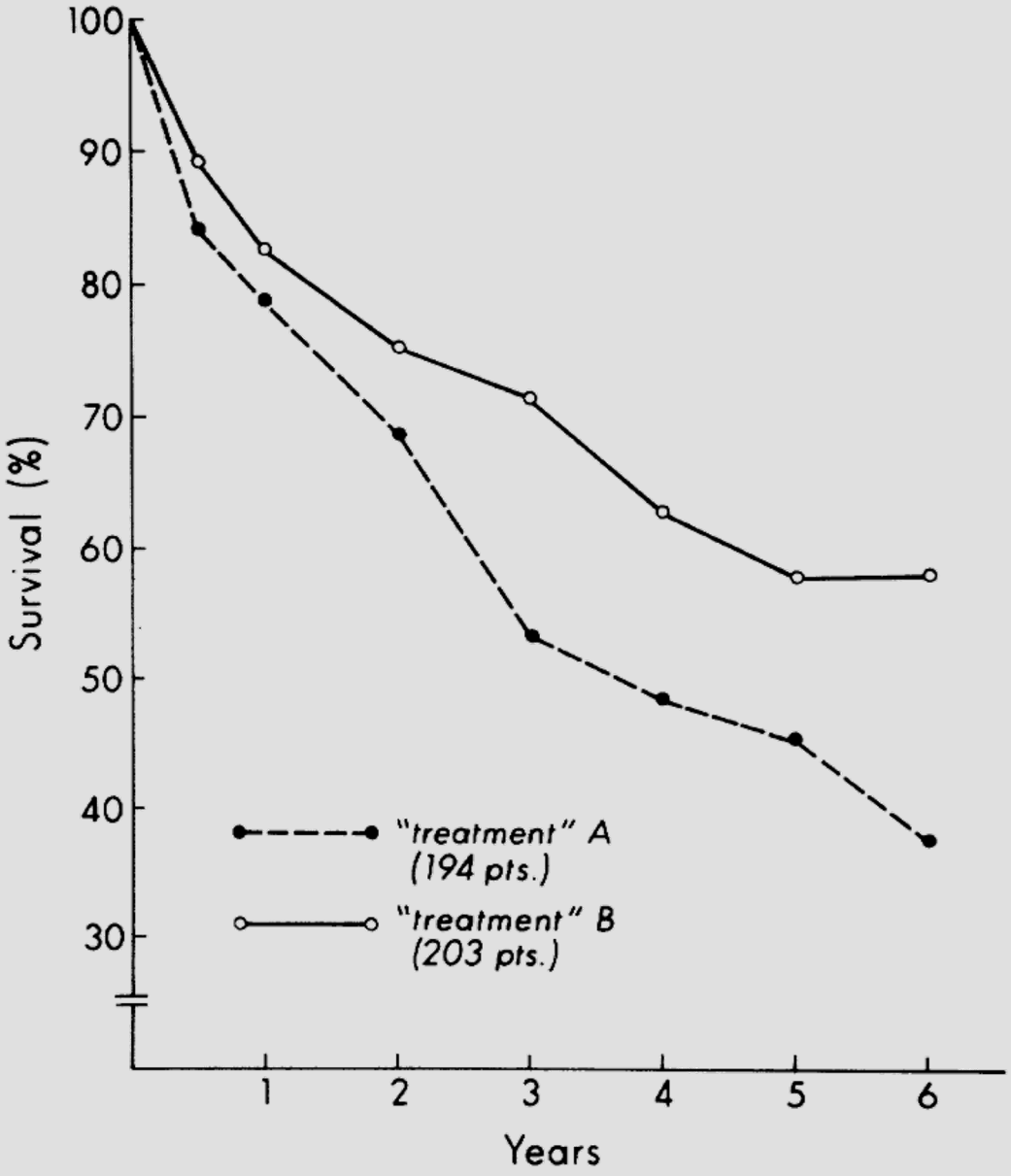
1073 patients with coronary heart disease were split in two treatment groups. We want to compare survival rates for two treatments.

Important variables influencing survival: the number of significantly diseased vessels (1,2,3) and left ventricular contraction pattern (normal/abnormal).

In one of six subgroups differences in survival rates were significant.

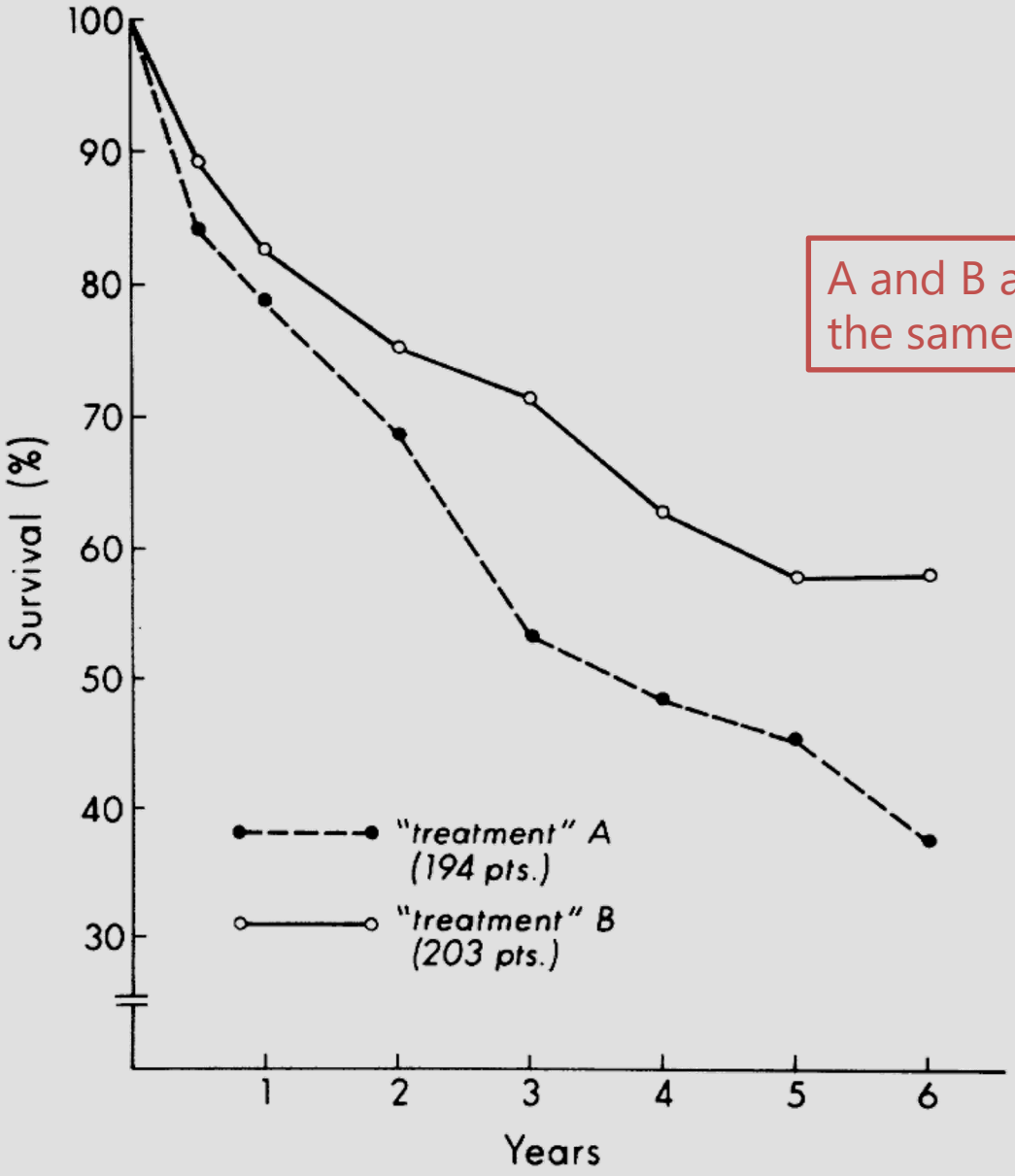
Lee et al, 1980

CHD treatments



Lee et al, 1980

CHD treatments



A and B are exactly the same treatment!

Lee et al, 1980

Caffeine and breast cancer

Caffeine consumption vs. breast cancer risk, around 50 subgroups. No overall risk increase, but:

- In women with benign breast disease, a borderline significant positive association with breast cancer risk was observed for the highest quintile of caffeine consumption and for coffee consumption ≥ 4 cups per day (both $p = 0.05$)
- Caffeine consumption was significantly positively associated with risk of estrogen receptor–negative and progesterone receptor–negative breast cancer and breast tumors larger than 2 cm (both $p = 0.02$)
- Inverse association between decaffeinated coffee consumption and risk in postmenopausal women that never used postmenopausal hormones ($p = 0.02$)

Ishitani, Lin, 2008

Subgroups on treatment outcome

Galen, 2nd century BC

«All who drink of this remedy recover in a short time, except those whom it does not help, who all die. Therefore, it is obvious that it fails only in incurable cases.»

wikipedia.org

Mutations

	Healthy controls (100)	Disease group (100)	p
Mutation	1	8	0.0349
Last name starts with a consonant	36	40	0.6622

Holm/Benjamini-Hochberg: nothing is rejected because $p_{(1)}$ is compared to $\frac{0.05}{2}$

Takeaways about subgroup analysis

- Always adjust for multiplicity
- For rigor, subgroups of interest need to be defined before seeing the data
- Less hypotheses, less pain

Usecases for linear regression

Linear regression model

$$y = x\beta + \varepsilon$$

- y – response variable
- $x = (1 \quad x_1 \quad \dots \quad x_k)$ – features
- $\beta = (\beta_0 \quad \beta_1 \quad \dots \quad \beta_k)^T$ – vector of coefficients such that $y \approx \beta x$

Uses of regression

$$y = x\beta + \varepsilon$$

- **Prediction**
- **Extrapolation**
- **Exploring associations**
- **Causal inference**

Gelman, Hill, Vehtari, Regression and other stories

Uses of regression

$$y = x\beta + \varepsilon$$

- **Prediction**
  Forecasting y on new data. Example: predicting sales of a product in different stores
- **Extrapolation**
- **Exploring associations**
- **Causal inference**

Uses of regression

$$y = x\beta + \varepsilon$$

- **Prediction**
  Forecasting y on new data. Example: predicting sales of a product in different stores
- **Extrapolation**
  Adjusting for known difference between sample and population. Example: debiasing a non-random survey
- **Exploring associations**
- **Causal inference**

Uses of regression

$$y = x\beta + \varepsilon$$

- **Prediction**
  Forecasting y on new data. Example: predicting sales of a product in different stores
- **Extrapolation**
  Adjusting for known difference between sample and population. Example: debiasing a non-random survey
- **Exploring associations**
  Summarizing how well variables predict the outcome. Example: studying association between height and income
- **Causal inference**

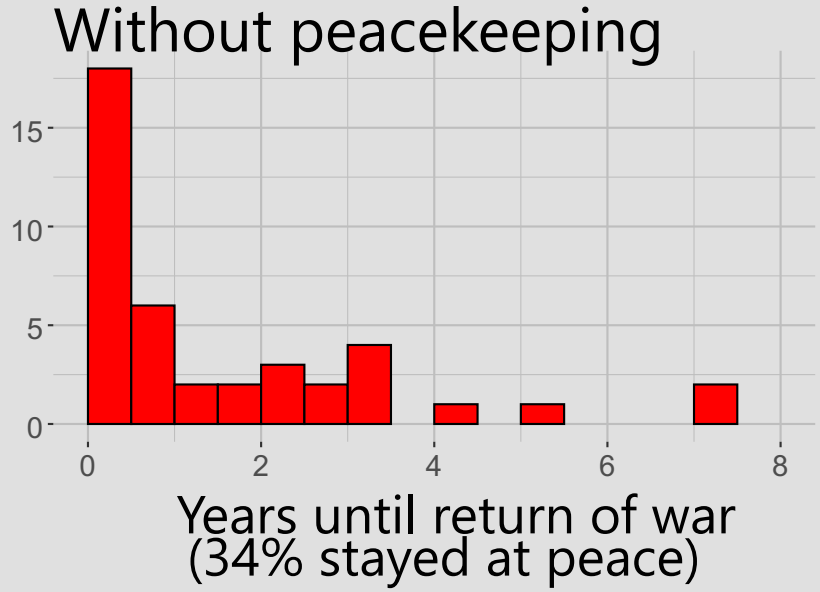
Uses of regression

$$y = x\beta + \varepsilon$$

- **Prediction**
  Forecasting y on new data. Example: predicting sales of a product in different stores
- **Extrapolation**
  Adjusting for known difference between sample and population. Example: debiasing a non-random survey
- **Exploring associations**
  Summarizing how well variables predict the outcome. Example: studying association between height and income
- **Causal inference**
  Estimating treatment effects. Example: quantifying the influence of air pollution level on disease incidence

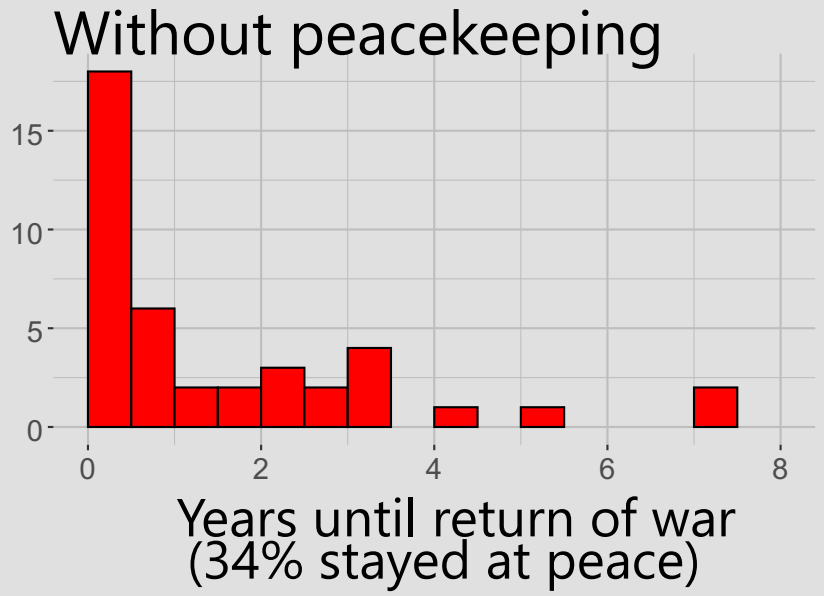
United Nations peacekeeping

Page Fortna peacekeeping study: data from countries involved in civil wars 1989-1999 with data collection stopped in 2004.



United Nations peacekeeping

Page Fortna peacekeeping study: data from countries involved in civil wars 1989-1999 with data collection stopped in 2004.



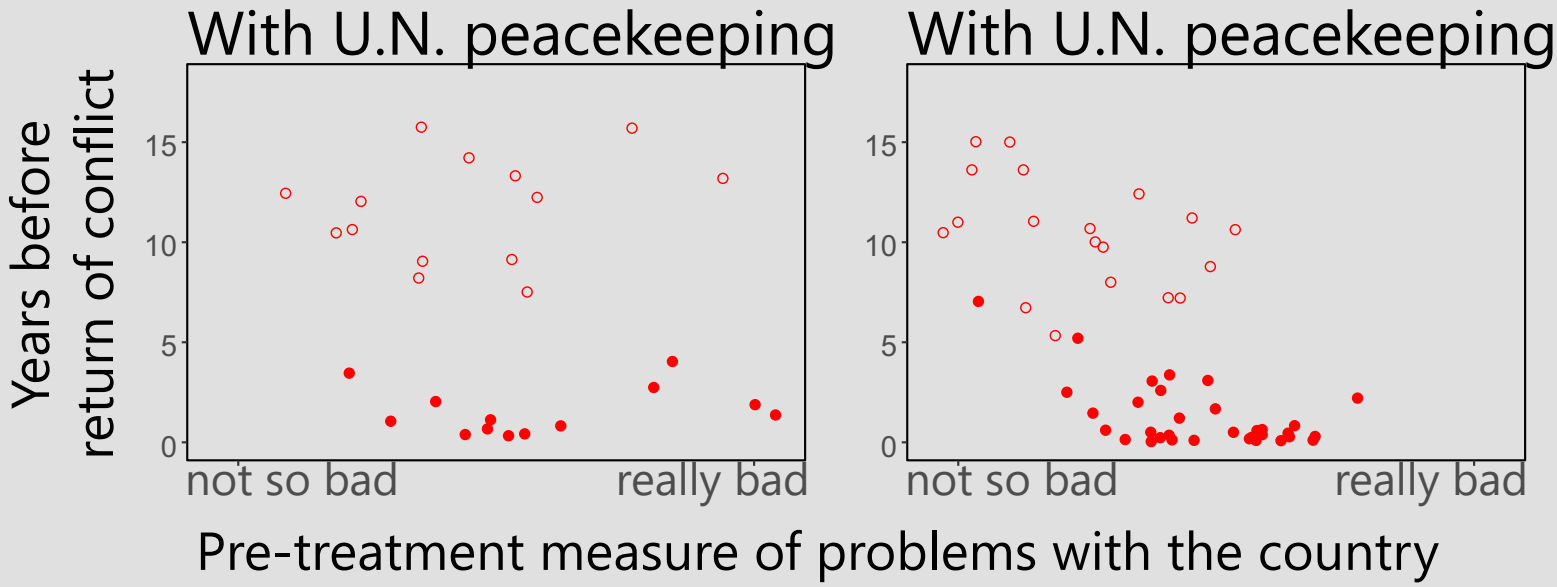
What if there's selection bias – maybe U.N. chooses the easy cases to get involved?

United Nations peacekeeping

How bad was the country before peacekeeping-or-not-peacekeeping decision had been made?

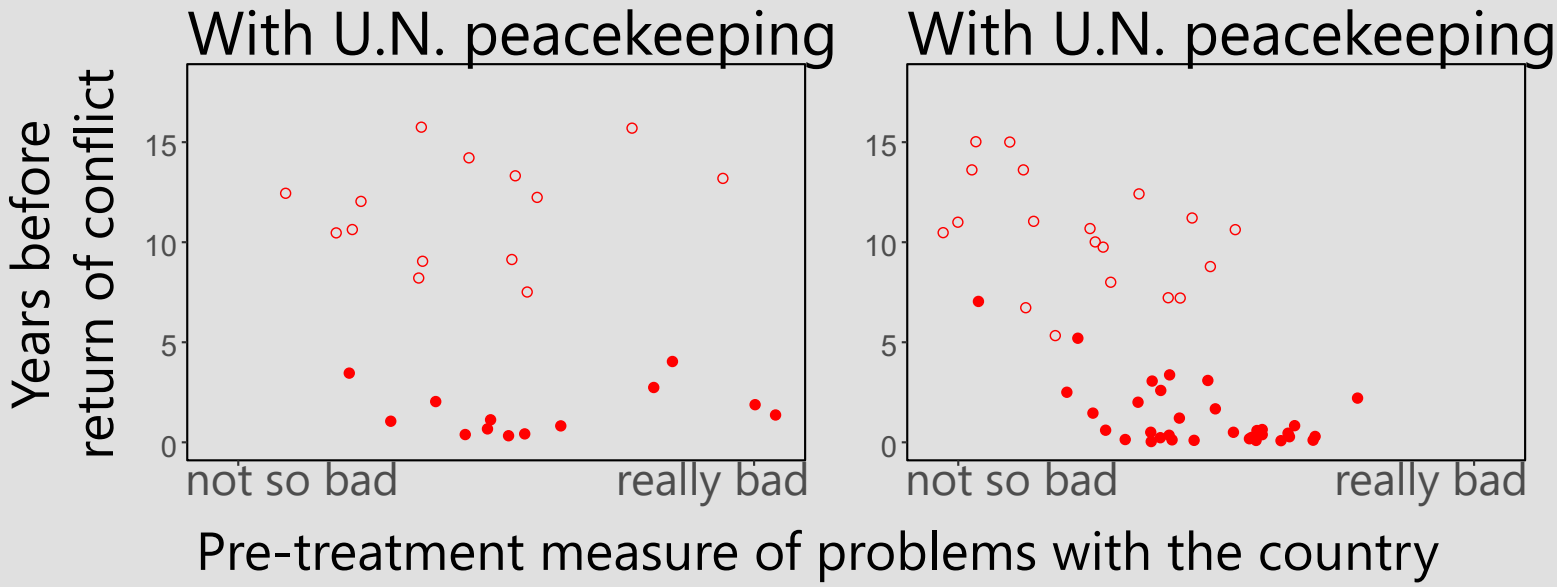
United Nations peacekeeping

How bad was the country before peacekeeping-or-not-peacekeeping decision had been made?



United Nations peacekeeping

How bad was the country before peacekeeping-or-not-peacekeeping decision had been made?



After adjusting for badness score, the estimate of the U.N. peacekeeping efficiency gets higher!

Estimation and properties

Linear regression model

$$y = x\beta + \varepsilon$$

- $1, \dots, n$ – objects in the sample

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Linear regression model

$$y = x\beta + \varepsilon$$

- $1, \dots, n$ – objects in the sample

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\hat{y} = X\hat{\beta}$$

Linear regression model

$$y = x\beta + \varepsilon$$

- $1, \dots, n$ – objects in the sample

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\hat{y} = X\hat{\beta}$$

- Least squares estimate:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2 = X(X^T X)^{-1} X^T y$$

Linear regression model

$$y = x\beta + \varepsilon$$

- $1, \dots, n$ – objects in the sample

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

$$\hat{y} = X\hat{\beta}$$

- Least squares estimate:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2 = X(X^T X)^{-1} X^T y$$

- With some math, it could be shown that

$$\hat{y} = x\hat{\beta} = \widehat{\mathbb{E}}(y|x)$$

Interpreting the coefficients

$$\mathbb{E}(y|x) = \beta_0 + \sum_{j=1}^{k} x_j \beta_j$$

β_j estimates how much does y change on average with unit change in x_j, holding all other features constant

Quality of the solution

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = ESS + RSS$$

Quality of the solution

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = ESS + RSS$$

Coefficient of determination:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Least squares properties

$$\hat{y} = X\hat{\beta}$$

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:

$$\mathbb{E}(\varepsilon|x) = 0$$

Least squares properties

$$\hat{y} = X\hat{\beta}$$

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:

$$\mathbb{E}(\varepsilon|x) = 0$$

$\Longrightarrow$ LSE $\hat{\beta}$ are unbiased and consistent

Least squares properties

$$\hat{y} = X\hat{\beta}$$

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:

$$\mathbb{E}(\varepsilon|x) = 0$$

5. Error is homoscedastic:

$$\mathbb{D}(\varepsilon|x) = \sigma^2$$

Least squares properties

$$\hat{y} = X\hat{\beta}$$

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:
$$\mathbb{E}(\varepsilon|x) = 0$$
5. Error is homoscedastic:
$$\mathbb{D}(\varepsilon|x) = \sigma^2$$

$\Longrightarrow$ LSE $\hat{\beta}$ have the smallest variance among all estimates of β that are linear on y

Least squares properties

$$\hat{y} = X\hat{\beta}$$

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:

$$\mathbb{E}(\varepsilon|x) = 0$$

5. Error is homoscedastic:

$$\mathbb{D}(\varepsilon|x) = \sigma^2$$

Gauss-Markov assumptions

$\Longrightarrow$ LSE $\hat{\beta}$ have the smallest variance among all estimates of β that are linear on y

Variance of LSE

In assumptions 1-5:

$$\mathbb{D}(\hat{\beta}_j) = \frac{\sigma^2}{TSS_j(1 - R_j^2)},$$

where $TSS_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$, R_j^2 – coefficient of determination when regressing x_j on all other features.

Variance of LSE

In assumptions 1-5:

$$\mathbb{D}(\hat{\beta}_j) = \frac{\sigma^2}{TSS_j(1 - R_j^2)},$$

where $TSS_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$, R_j^2 – coefficient of determination when regressing x_j on all other features.

Variance of $\hat{\beta}_j$ grows:

- with noise level
- if x_j gets closer to a constant
- if x_j gets closer to a linear combination of other features

Variance of LSE

$R_j^2 < 1$ under assumption 3, but it still could be ≈ 1

In matrix notation:

$$\mathbb{D}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

If columns of X are almost linearly dependent, then inverting $X^T X$ is unstable and the variance of $\hat{\beta}$ is large – **multicollinearity**

Least squares properties

$$\hat{y} = X\hat{\beta}$$

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:
$$\mathbb{E}(\varepsilon|x) = 0$$
5. Error is homoscedastic:
$$\mathbb{D}(\varepsilon|x) = \sigma^2$$
6. Error is Gaussian:
$$\varepsilon|x \sim N(0, \sigma^2)$$

Least squares properties

$$\hat{y} = X\hat{\beta}$$

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:
$$\mathbb{E}(\varepsilon|x) = 0$$
5. Error is homoscedastic:
$$\mathbb{D}(\varepsilon|x) = \sigma^2$$
6. Error is Gaussian:
$$\varepsilon|x \sim N(0, \sigma^2)$$

$\Longrightarrow$ LSE $\hat{\beta}$ are MLE

LSE when they are MLE

- Have the smallest variance among all unbiased estimates of β
- $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$
- $\hat{\sigma}^2 = \frac{1}{n-k-1} RSS$ – unbiased estimate of σ^2, and

$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-k-1}$$

- $\forall c \in \mathbb{R}^{k+1}$:

$$\frac{c^T(\beta - \hat{\beta})}{\hat{\sigma}\sqrt{c^T (X^T X)^{-1} c}} \sim St(n-k-1)$$

Under assumptions 1-6 we could test hypotheses and build confidence intervals for coefficients β_j!

Takeaways about LSE

- Under certain assumptions, LSE regression solution has some nice statistical properties
- To have a good estimate of a feature coefficient, that feature needs to be good, and noise level needs to be low
- If the noise is Gaussian, we know how different regression products are distributed

Inference

Confidence intervals

$\forall c \in \mathbb{R}^{k+1}$:

$$\frac{c^T(\beta - \hat{\beta})}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim St(n-k-1)$$

Confidence intervals

$\forall c \in \mathbb{R}^{k+1}$:

$$\frac{c^T(\beta - \hat{\beta})}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim St(n-k-1)$$

- Take $c = (c_0, c_1, \dots, c_k), c_i = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

$\Longrightarrow 100(1-\alpha)\%$ confidence interval for β_j:

$$\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$$

Confidence intervals

$\forall c \in \mathbb{R}^{k+1}$:

$$\frac{c^T(\beta - \hat{\beta})}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim St(n-k-1)$$

- For an object with features x_0 take $c = (1, x_{01}, \dots, x_{0k})$;

$\Longrightarrow 100(1-\alpha)\%$ confidence interval for $\mathbb{E}(y|x = x_0)$:

$$x_0^T\hat{\beta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}$$

Confidence intervals

$\forall c \in \mathbb{R}^{k+1}$:

$$\frac{c^T(\beta - \hat{\beta})}{\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}} \sim St(n-k-1)$$

- For an object with features x_0 take $c = (1, x_{01}, \dots, x_{0k})$;

$\Longrightarrow 100(1-\alpha)\%$ confidence interval for $\mathbb{E}(y|x = x_0)$:

$$x_0^T\hat{\beta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}$$

$\Longrightarrow 100(1-\alpha)\%$ prediction interval for $y(x_0)$:

$$x_0^T\hat{\beta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$$

T test for a coefficient

null hypothesis: $H_0: \beta_j = c$

alternative hypothesis: $H_1: \beta_j <\neq> c$

statistic: $$T = \frac{\hat{\beta}_j - c}{\sqrt{\frac{RSS}{n-k-1}(X^T X)^{-1}_{jj}}}$$

null distribution: $St(n-k-1)$

T test for a coefficient

Example: 12 subjects, x – composite reaction time test, y – flight simulator score. x is much easier and cheaper to measure, can we use it to predict y?

$$y = \beta_0 + x\beta_1 + \varepsilon$$

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0 \implies p = 2.2 \times 10^{-5}$, x is useful for predicting y.

F test for several coefficients

$$\underset{k+1}{x} = \underset{k+1}{(} \underset{k+1-k_1}{x_1}, \underset{k_1}{x_2} \underset{}{)}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \begin{matrix} k+1-k_1 \\ k_1 \end{matrix}$$

null hypothesis: $H_0: \beta_2 = 0$

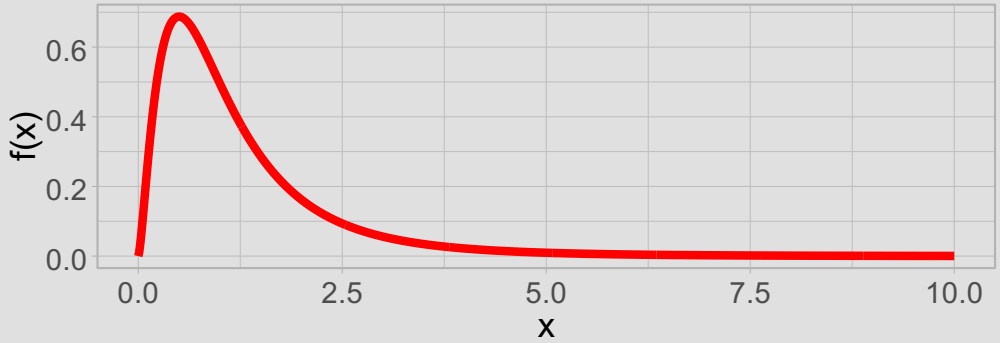
alternative hypothesis: $H_1: \beta_2 \neq 0$

RSS_r – RSS when regressing y on x_1

RSS_{ur} – RSS when regressing y on x

statistic: $$F = \frac{(RSS_r - RSS_{ur})/k_1}{RSS_{ur}/(n-k-1)}$$

null distribution: $F(k_1, n-k-1)$



F test for several coefficients

Example: data on 1191 children, regression model:

$$weight = \beta_0 + cigs\beta_1 + parity\beta_2 + inc\beta_3 + med\beta_4 + fed\beta_5 + \varepsilon$$

$weight$ – birthweight
$cigs$ – average number of cigarettes smoked per day of pregnancy
$parity$ – order number of the child for their mother
inc – mean monthly income
med – duration of education for mother
fed – for father

Are the education variables good predictors of birthweight?

F test for several coefficients

Example: data on 1191 children, regression model:

$$weight = \beta_0 + cigs\beta_1 + parity\beta_2 + inc\beta_3 + med\beta_4 + fed\beta_5 + \varepsilon$$

Are the education variables good predictors of birthweight?

$H_0: \beta_4 = \beta_5 = 0$
$H_1: H_0$ is false $\Longrightarrow p = 0.2421$, we cannot reject the hypothesis that med and fed are useless for predicting y given remaining variables.

F versus T test

- When $k_1 = 1$, F test is equivalent to T test with two-sided alternative

F versus T test

- When $k_1 = 1$, F test is equivalent to T test with two-sided alternative
- Sometimes F rejects the null, while T does not reject the null for any of the components of x_2. Possible explanations:
  - Separately, each component of x_2 is not a good enough predictor, but together they work
  - X_2 has multicollinearity

F versus T test

- When $k_1 = 1$, F test is equivalent to T test with two-sided alternative
- Sometimes F rejects the null, while T does not reject the null for any of the components of x_2. Possible explanations:
  - Separately, each component of x_2 is not a good enough predictor, but together they work
  - X_2 has multicollinearity
- Sometimes F does not reject the null, while T rejects the null for some of the components of x_2. Possible explanations:
  - Useless features in x_2 are masking the effect of useful ones
  - Significance of some coefficients in β_2 is a result of multiple hypotheses testing

Testing the assumptions

LSE assumptions

1. Linearity of the response on features (model is true)
2. Simple random sampling
3. Matrix X is full rank
4. Error is random:

$$\mathbb{E}(\varepsilon|x) = 0$$

5. Error is homoscedastic:

$$\mathbb{D}(\varepsilon|x) = \sigma^2$$

6. Error is Gaussian:

$$\varepsilon|x \sim N(0, \sigma^2)$$

1. Linearity of the response

$$y = x\beta + \varepsilon$$

Never exactly true – remember, all models are false!
To test whether there are large deviations from linearity, we need to look at residuals:

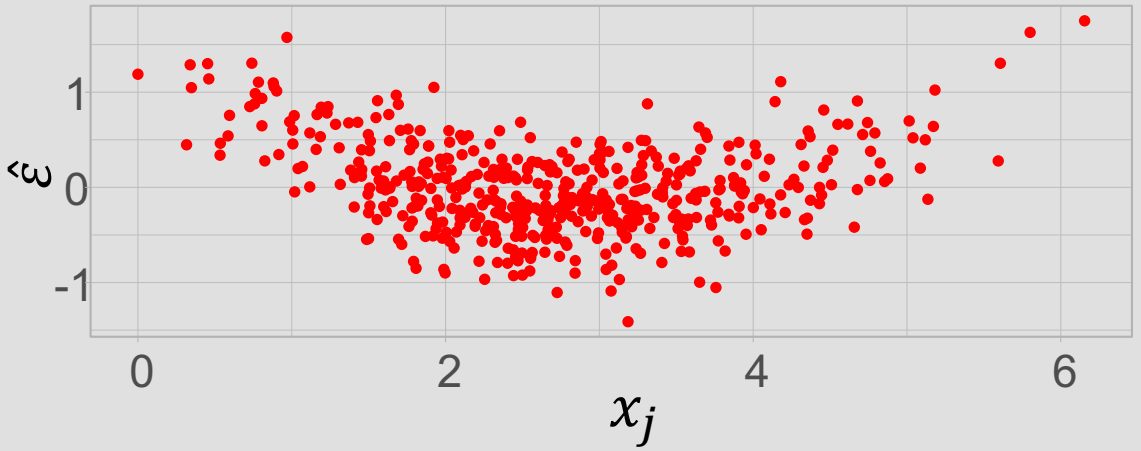
$$\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1, \dots, n$$

1. Linearity of the response

$$y = x\beta + \varepsilon$$

Never exactly true – remember, all models are false!
To test whether there are large deviations from linearity, we need to look at residuals:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1, \dots, n$$



Makes sense to add x_j^2 as a feature

2. Simple random sampling

The observations (x_i, y_i) are i.i.d.

- If the observations are dependent, the variance is underestimated, and tests do not work properly
- An easy way to violate this assumption is to filter the sample by a feature z; it is only allowed when

$$\mathbb{E}(y|x) = \mathbb{E}(y|x, z)$$

3. Matrix X is full rank

$$\operatorname{rank} X = k + 1$$

- For linearly dependent features the variance of coefficients would be infinite
- One-hot encoding is not allowed! But dummy encoding is.

3. Matrix X is full rank

$$\operatorname{rank} X = k + 1$$

- For linearly dependent features the variance of coefficients would be infinite
- One-hot encoding is not allowed! But dummy encoding is.

y – salary, x – position:

Position	x_1	x_2
worker	0	0
engineer	1	0
manager	0	1

$$\Rightarrow y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \varepsilon$$

β_1 and β_2 estimate the average difference in salary levels between engineer and worker, manager and worker.

4. Error is random

$$\mathbb{E}(\varepsilon|x) = 0$$

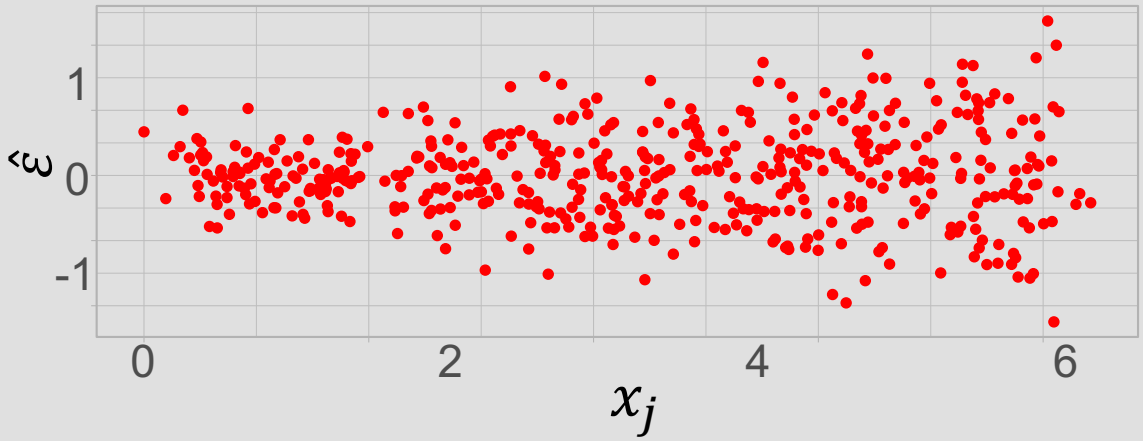
In theory, H_0: $\mathbb{E}(\varepsilon|x) = 0$ could be tested on residuals with one sample T test, but in practice mean residuals are always 0

5. Error is homoscedastic

$$\mathbb{D}(\varepsilon|x) = \sigma^2$$

Ways to test:

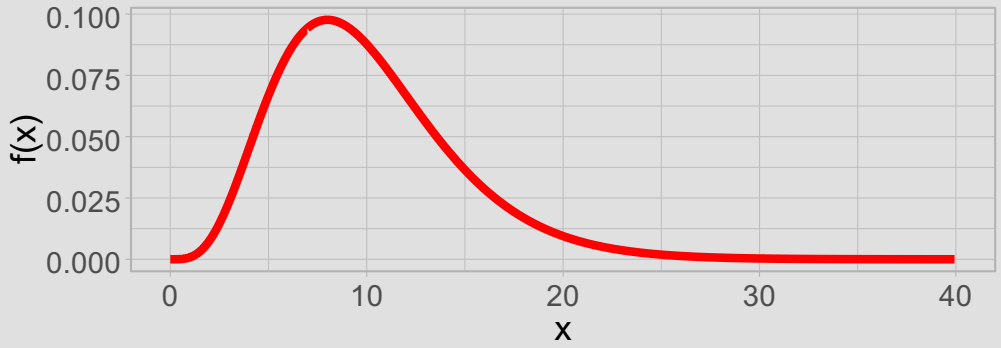
- visual analysis:



- Breusch–Pagan test

Breusch–Pagan test

null hypothesis:	$H_0: \mathbb{D}(\varepsilon \mid x) = \sigma^2$
alternative hypothesis:	$H_1: H_0$ is false
statistic:	$LM = nR^2_{\hat{\varepsilon}^2}$, $R^2_{\hat{\varepsilon}^2}$ – coefficient of determination when regressing $\hat{\varepsilon}^2$ on x
null distribution:	χ^2_k



6. Error is Gaussian

$$\varepsilon|x \sim N(0, \sigma^2)$$

Tested on residuals:

- visually with q-q plot
- formally with Shapiro-Wilk test

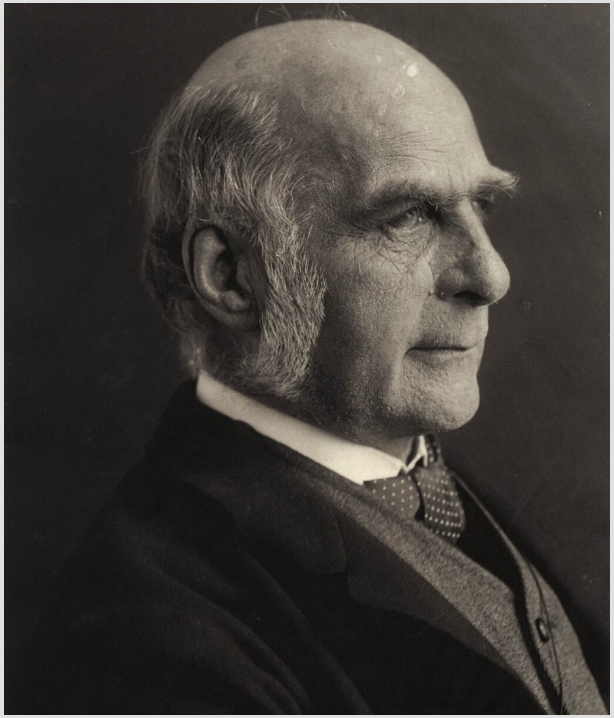
Takeaways

- Linear regression assumptions are (mostly) testable
- Remember why each assumption is needed! E.g., if you do not plan to do inference, normality is not necessary.

Regression to the mean

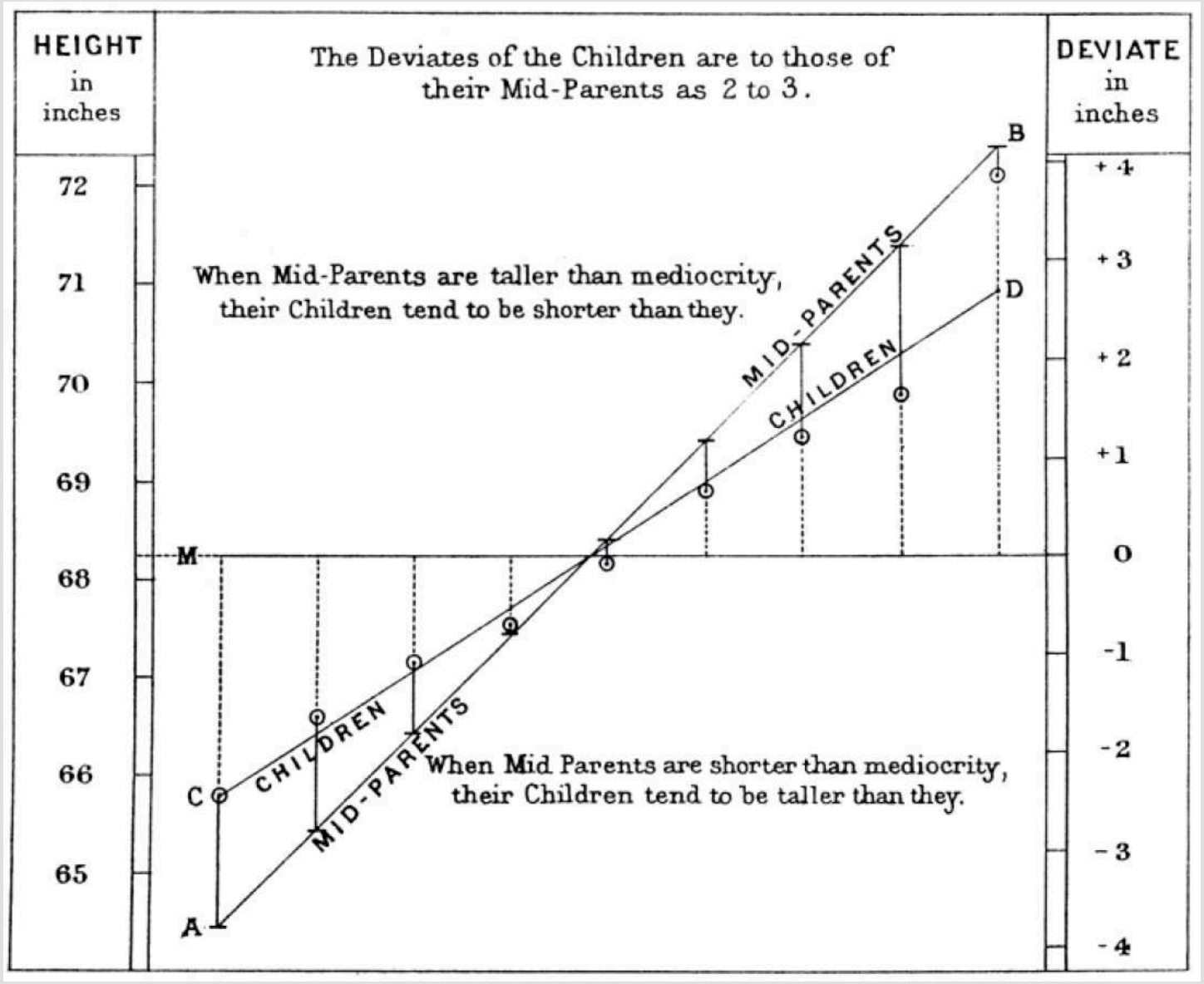
Francis Galton

- Regression, correlation, quartiles, median, standard deviation
- Dactyloscopy
- Theory of anticyclones
- Number form synesthesia
- Ultrasonic dog whistle
- Eugenics



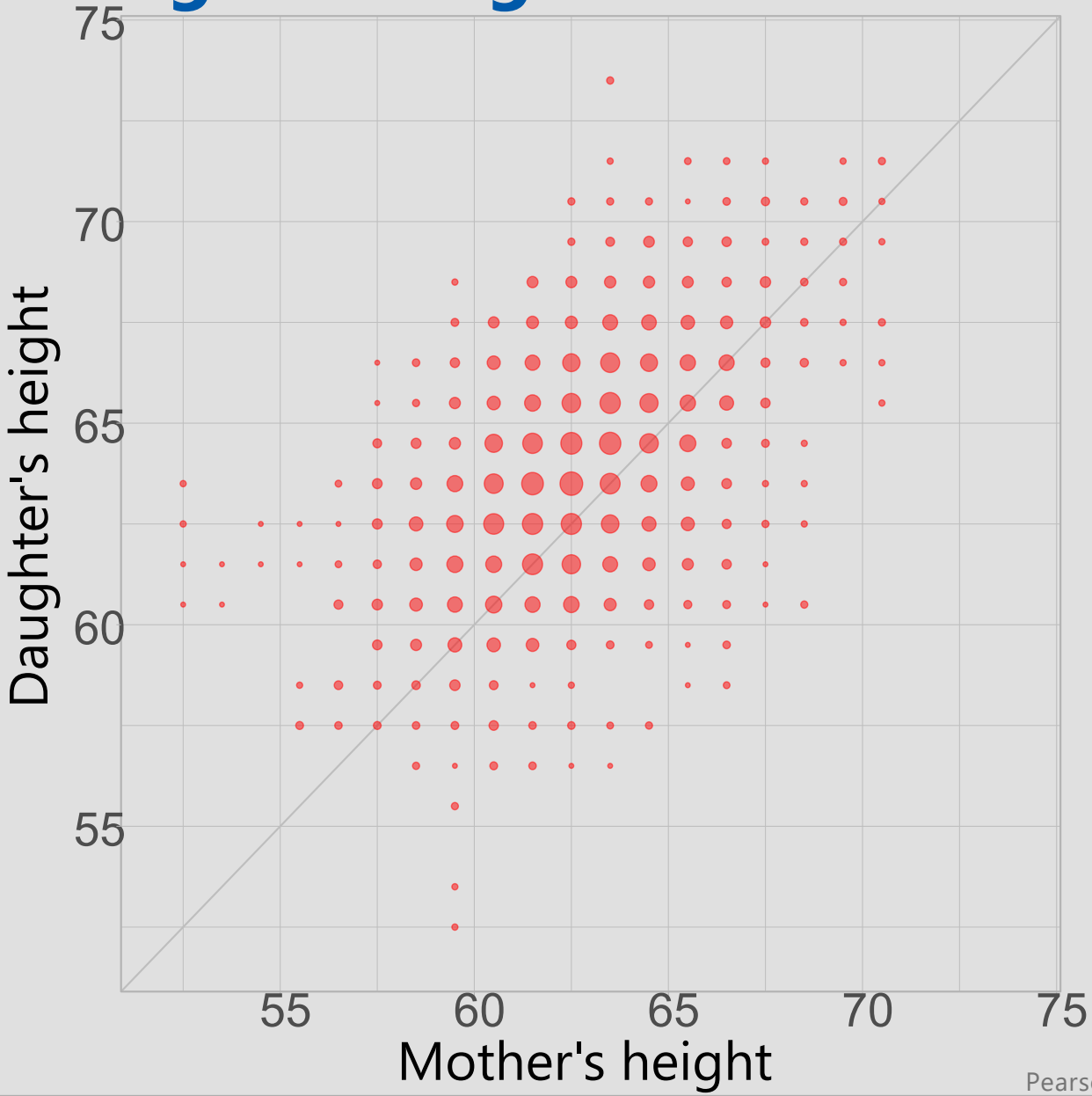
wikipedia.org

The origin of regression



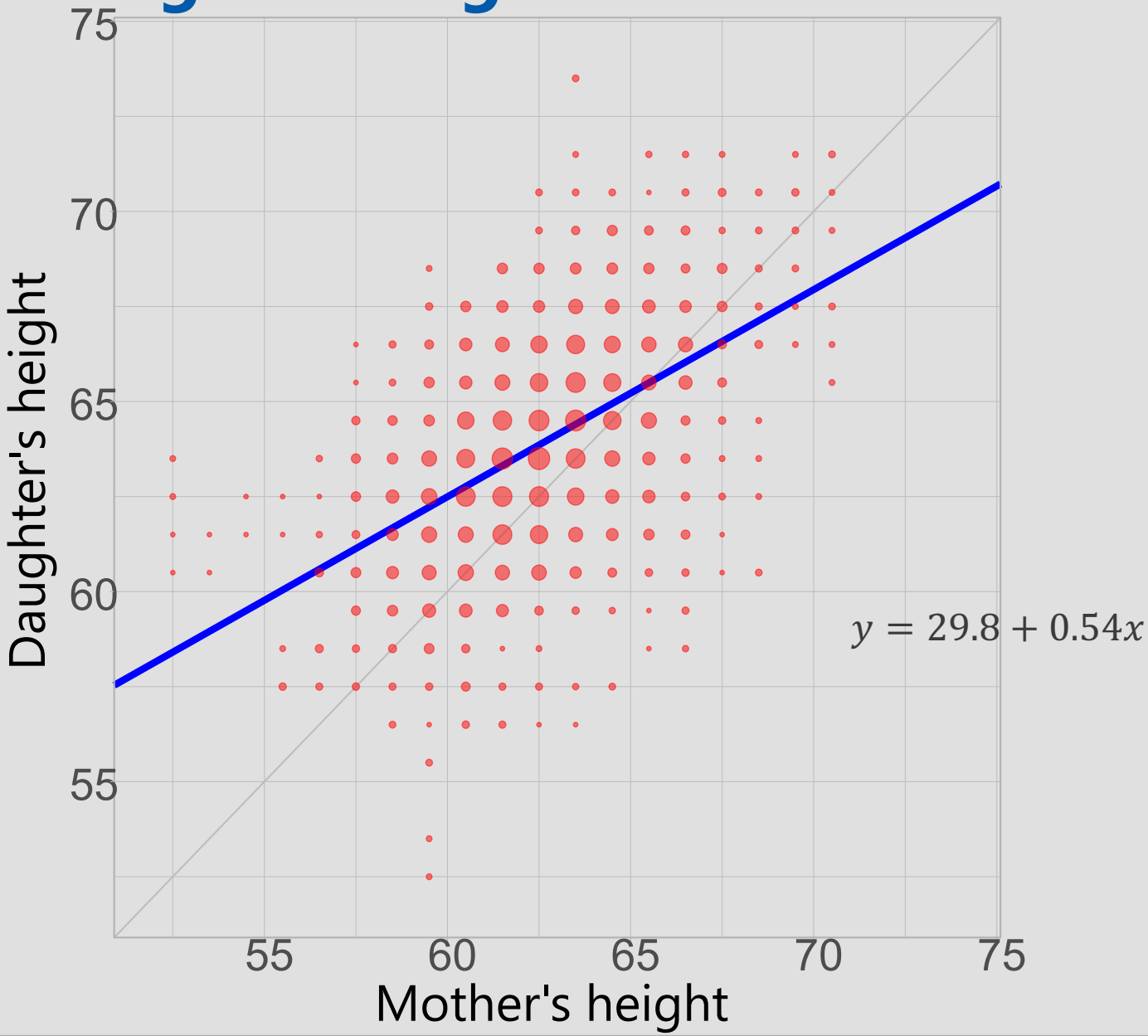
Galton, Regression towards mediocrity in hereditary stature, 1886

Mother-daughter heights

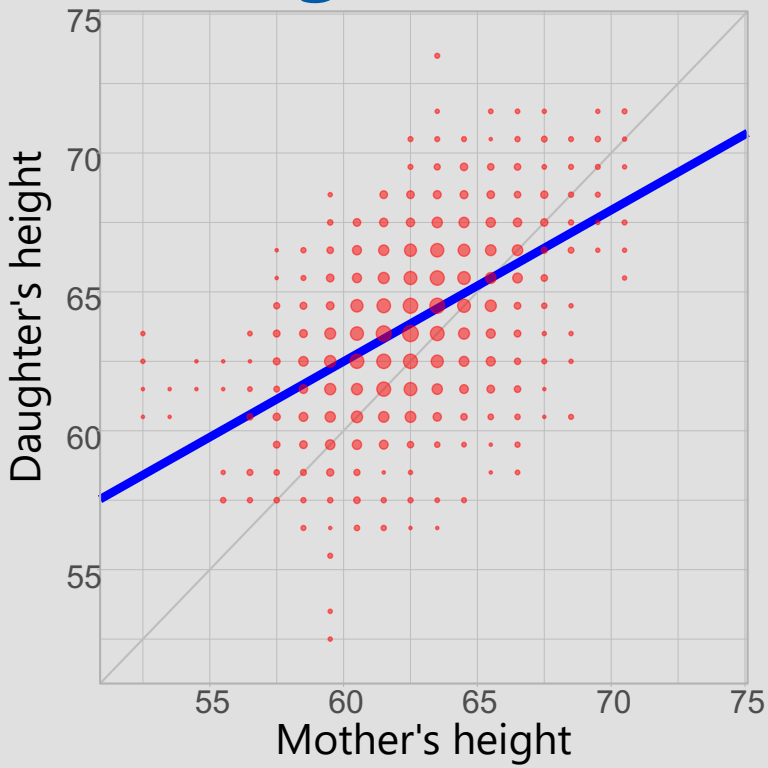


Pearson, Lee, 1903

Mother-daughter heights



Mother-daughter heights

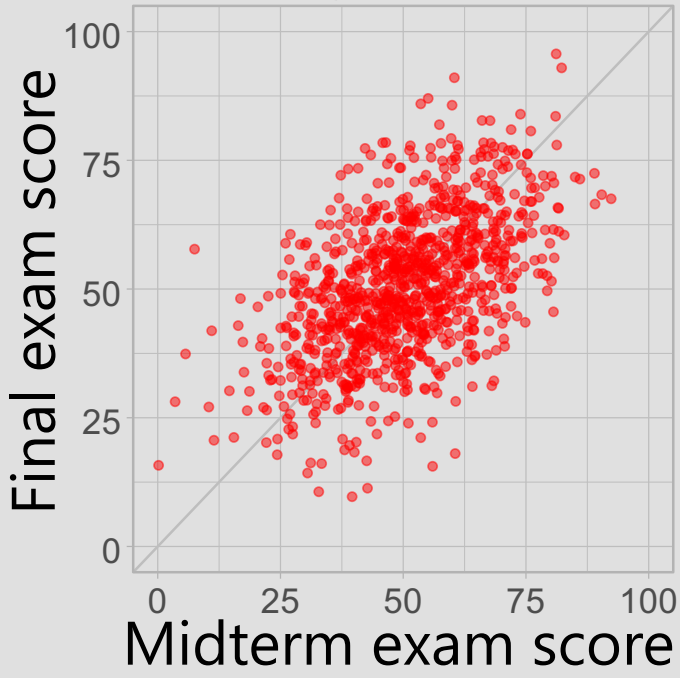


- Woman's *predicted* height is closer to average than her mother's; not her *actual* height!
- A random error – unpredictable component – is making sure of that

Another example

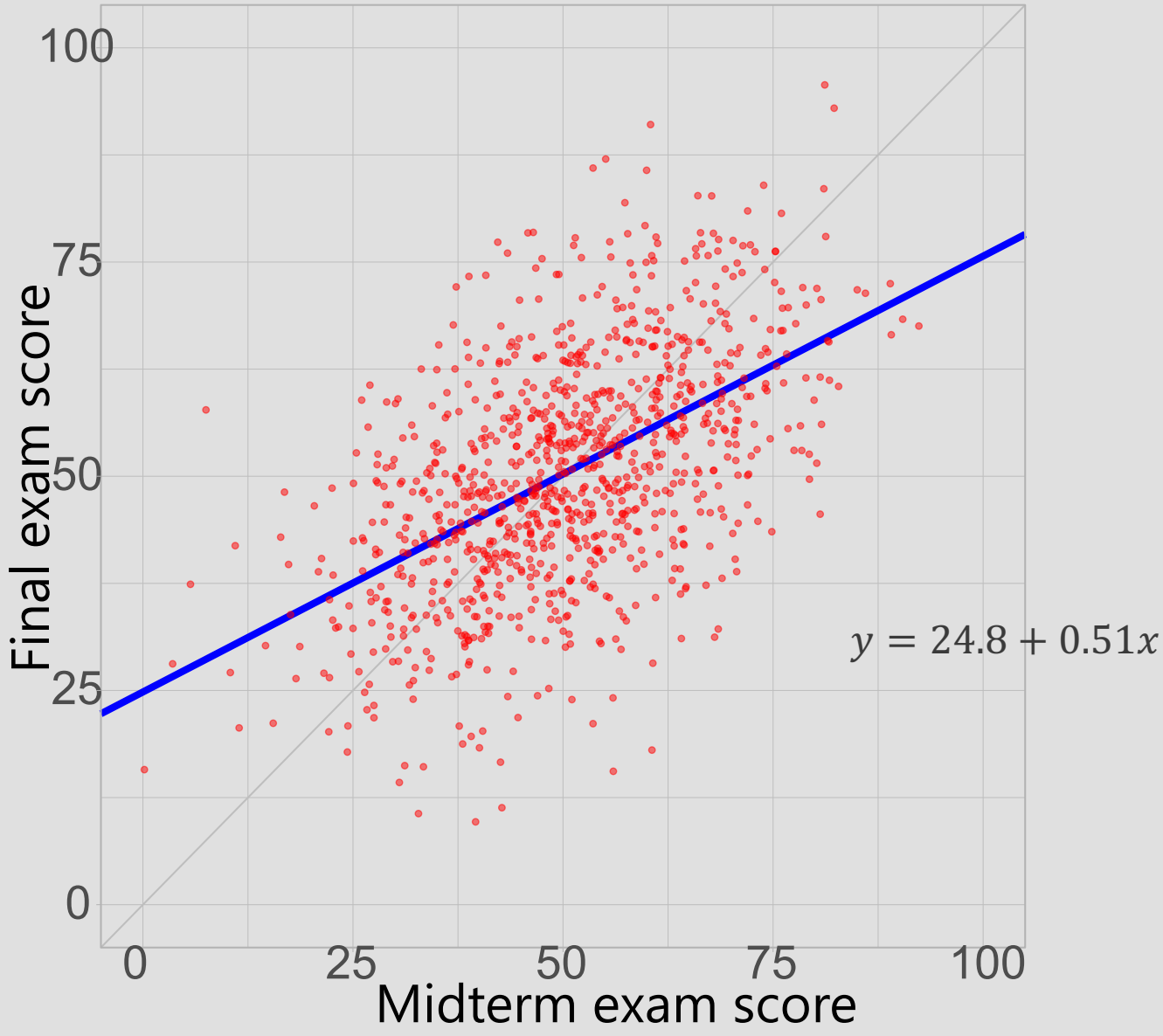
For 1000 students we are going to simulate their mid-term and final exam scores:

- Each student is assumed to have a true ability – we'll draw it from $N(50, 100)$
- In both exams, each student's score would consist of two components – a true ability plus random noise (from $N(0, 100)$)



Gelman, Hill, Vehtari, Regression and other stories

Another example



Another example

Naïve interpretation of the data: better scoring students are getting lazy on the final; worse scoring students are studying harder – **regression fallacy**

Another example

Naïve interpretation of the data: better scoring students are getting lazy on the final; worse scoring students are studying harder – **regression fallacy**

True interpretation: students who scored very well on midterm have high ability, but they also got lucky! We expect them to score lower on the final.

Yet another example

The instructors in a flight school adopted a policy of consistent positive reinforcement recommended by psychologists. They verbally reinforced each successful execution of a flight maneuver. After some experience with training approach, the instructors claimed that contrary to psychological doctrine, high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try.

Tversky, Kahneman, 1973

Many more examples

- When his pain got worse, he went to a doctor, after which the pain subsided a little. Therefore, he benefited from the doctor's treatment.

- The student did exceptionally poorly last semester, so I punished him. He did much better this semester. Clearly, punishment is effective in improving students' grades.

- The frequency of accidents on a road fell after a speed camera was installed. Therefore, the speed camera has improved road safety.

Don’t fall for regression fallacy!

BONUS: multiple hypotheses testing in linear regression