Philosophy of Statistics

Daniel Kenneally

May 19, 2022

Science to Statistics

"Statistics is applied philosophy of Science" - *Kempthorne (1976)*

The Scientific Method

The Scientific Method

- A scientist creates a hypothesis.
- A **hypothesis** is a supposition or proposed explanation of a phenomenon.
- They observe the world and then they deduce from their observation whether the hypothesis was falsified (proved to be false).

The Scientific Method

- A scientist creates a hypothesis.
- A **hypothesis** is a supposition or proposed explanation of a phenomenon.
- They observe the world and then they deduce from their observation whether the hypothesis was falsified (proved to be false).

Example:

- Newton supposes an explanation of how the planets move.
- The planets arrive at their predicted destination.
- He cannot reject his hypothesis.

Definitions

H_0 : The null hypothesis

H_a : The alternative hypothesis

$\alpha = \text{P(false rejection (Type-I Error))}$

$\beta = \text{P(false acceptance (Type-II Error))}$

Definitions

H_0 : The null hypothesis

H_a : The alternative hypothesis

$\alpha = \text{P(false rejection (Type-I Error))}$

$\beta = \text{P(false acceptance (Type-II Error))}$

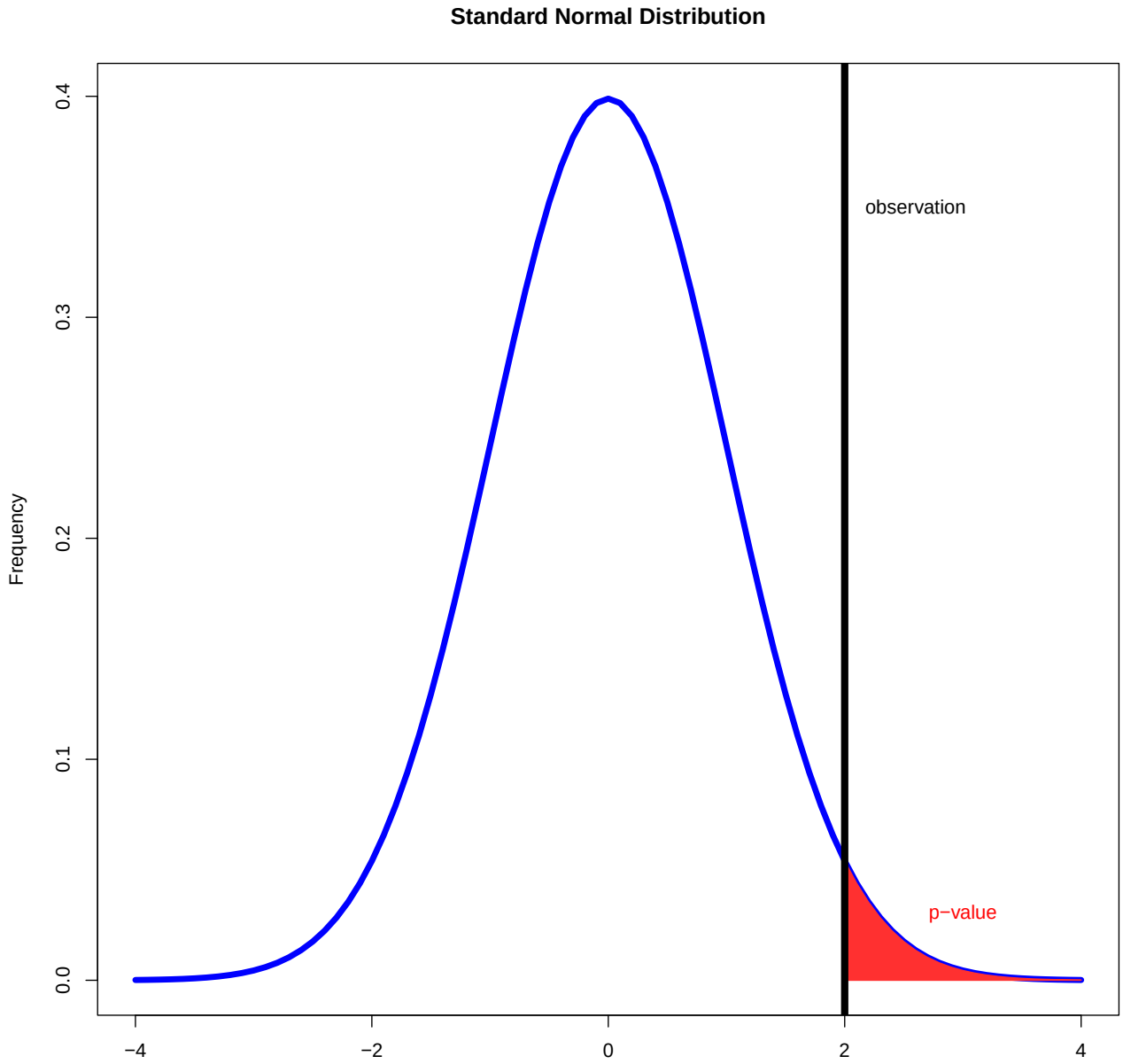
Definition

The **significance level** for the test is used to assess the **p-value** which is defined as follows:

$$\text{p-value} = P(\text{Data} \mid H_0)$$

i.e. the probability that the test statistic be at least as large as the value obtained from the data, given that the null hypothesis H_0 is true.

p-value


Standard Normal Distribution
Frequency
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
observation
p−value

Main Philosophies of Hypothesis Testing

Fisher's Null Hypothesis Testing

1. Set up a statistical null hypothesis.
2. Report the exact level of significance (e.g. p-value $= 0.051$ or p-value $= 0.049$).
3. Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses.
4. *Fisher (1955)*[1] "We have the duty of [...] communicating our conclusions, in intelligible form, in recognition of the right of *other* free minds to utilize them in making their own decisions".

Fisher's Null Hypothesis Testing

1. Set up a statistical null hypothesis.
2. Report the exact level of significance (e.g. p-value $= 0.051$ or p-value $= 0.049$).
3. Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses.
4. *Fisher (1955)*[1] "We have the duty of [...] communicating our conclusions, in intelligible form, in recognition of the right of *other* free minds to utilize them in making their own decisions".

Limits

- Analysis relies on subjective interpretation.

Neyman and Pearson Decision Theory

1. Set up two statistical hypotheses, H_0 and H_a.
2. Decide about α, β, and sample size before the experiment, based on a cost-benefit analysis. These define a rejection region for each hypothesis.
3. If the data falls into the rejection region of H_0, accept H_a; otherwise accept H_0.

Neyman and Pearson Decision Theory

1. Set up two statistical hypotheses, H_0 and H_a.
2. Decide about α, β, and sample size before the experiment, based on a cost-benefit analysis. These define a rejection region for each hypothesis.
3. If the data falls into the rejection region of H_0, accept H_a; otherwise accept H_0.

Limits

- Subjective choice of α and β.
- Limited to situations where there is a meaningful cost-benefit trade-off for choosing alpha and beta.
- Limited to situations where you have a disjunction of hypotheses (e.g., either $\mu_1 = 8$ or $\mu_2 = 10$ is true)

Philosophical Disagreements

Philosophical Disagreements

Philosophy:

- Fisher method aims to produce scientific truth.
- Neyman and Pearson method is an acceptance procedure.

Philosophical Disagreements

Philosophy:

- Fisher method aims to produce scientific truth.
- Neyman and Pearson method is an acceptance procedure.

Disagreements:

- For Fisher, Neyman's method is unscientific and only useful in limited cases within industry.
- For Neyman, Fisher's method requires a significance threshold to be useful and does not account for consequences of mistakes.

Current Practices

Current Practices

Gigerenzer (2004)[2]:

1. Set up a statistical null hypothesis.
2. Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis.
3. Report the result as $p < 0.05$, $p < 0.01$, or $p < 0.001$ (whichever comes next to the obtained p-value).
4. Always perform this procedure.

Misusing Statistics

Misusing Statistics

- Statistical metrics are stochastic/random variables.
- Statistical metrics are computations on a sample of observations.
- *Neyman (1957)*[3]: "these observations are random variables, the results of the computations are also random."
- Redraw samples, calculate the metric and take the upper/lower bound.
- The distribution of the maximum/minimum is different to the underlying distribution.

Selection of Best p-value

Selection of Best p-value

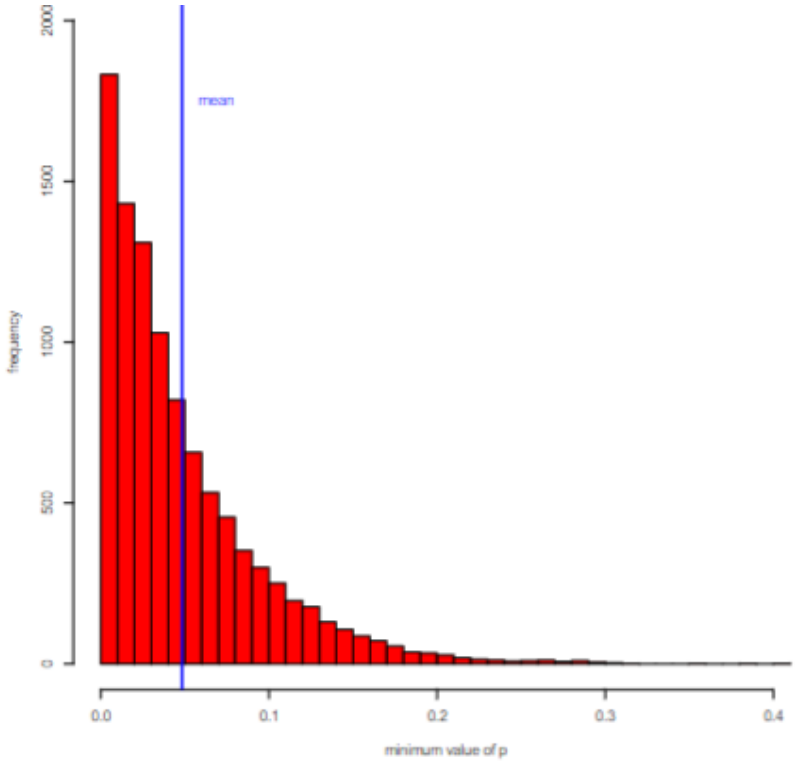
Taleb (2018)[4] illustrates;
Monte Carlo (M = 1000):

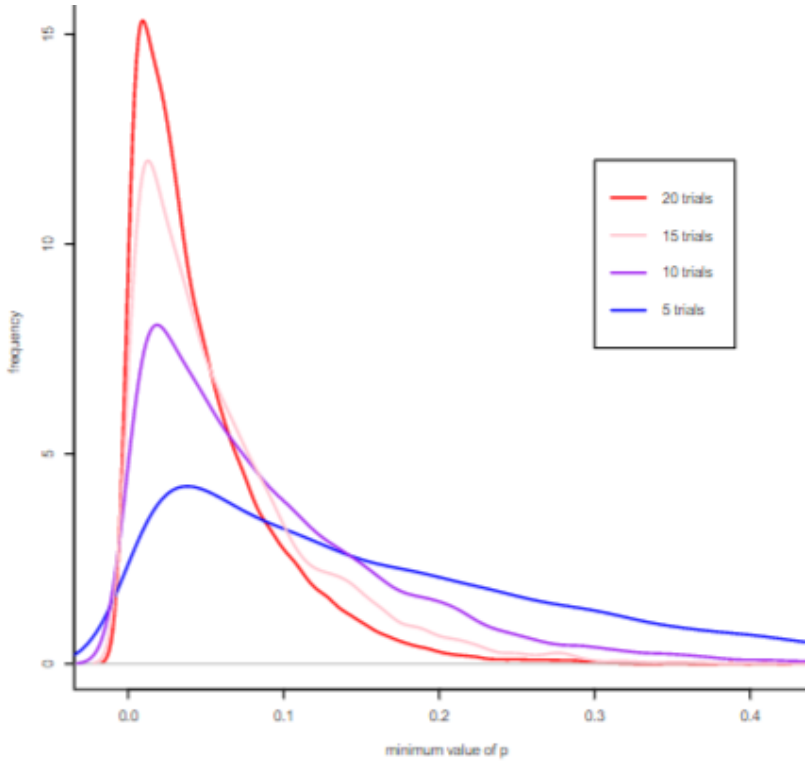
1. A trial consists of selecting two sets (X, Y), of 100 samples from a *N(0,1)* distribution.
2. Apply linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0,1)$$

3. H_0: $\beta_1 = 0$
4. Complete t trials.
5. Select the trial with the minimum p-value for publication.

Distribution of minimum p-value





Trials	5	10	15	20
Mean	0.1676	0.0909	0.0627	0.0483
Proportion	0.2316	0.4039	0.5365	0.6423

Table 4.1: Mean and proportion of statistically significant p-values

Repeated Sampling

Repeated Sampling

Monte Carlo (M = 1000):

1. Draw two sets (X, Y), of 20 samples from a standard normal distribution, *N(0,1)*.
2. Apply linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0,1)$$

3. H_0: $\beta_1 = 0$

Repeated Sampling

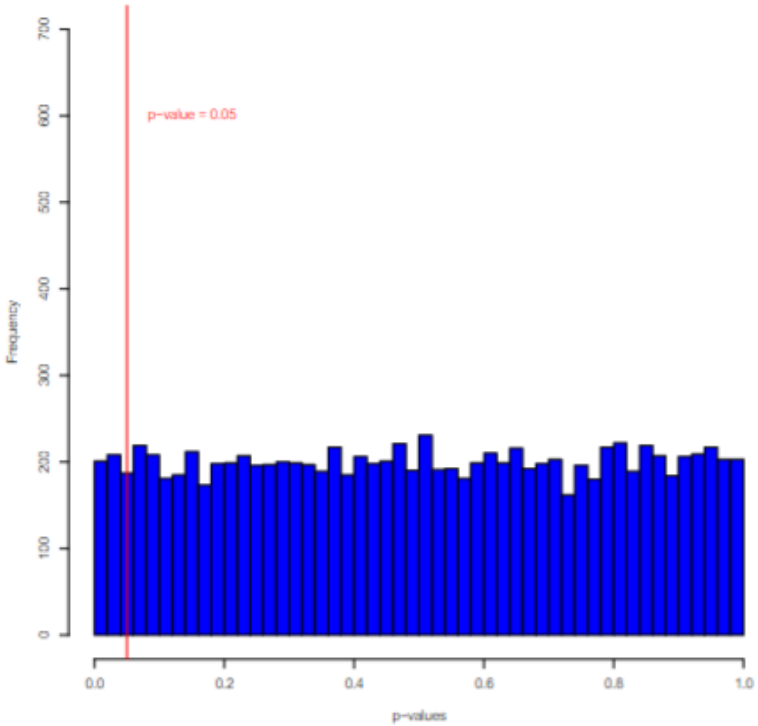
Monte Carlo (M = 1000):

1. Draw two sets (X, Y), of 20 samples from a standard normal distribution, *N(0,1)*.
2. Apply linear regression:

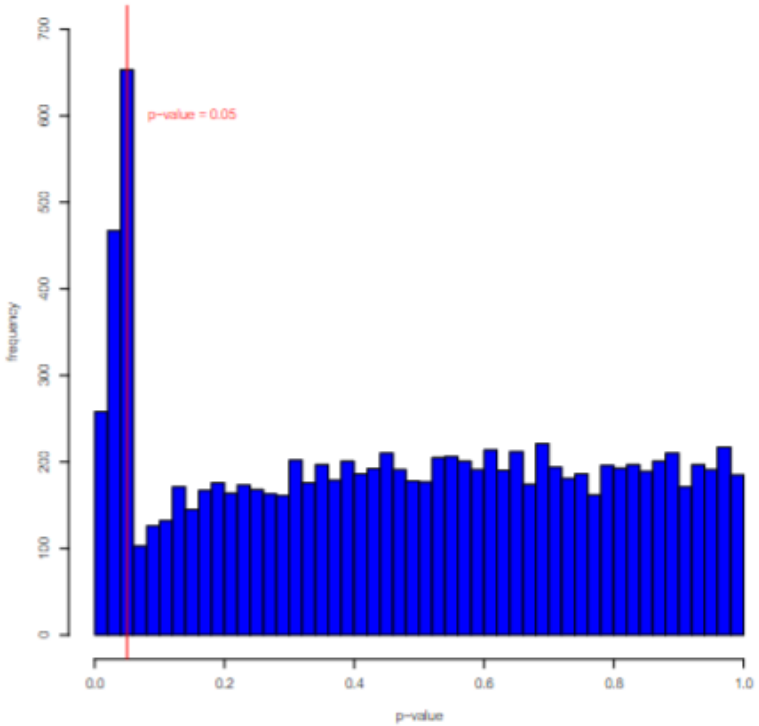
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0,1)$$

3. H_0: $\beta_1 = 0$
4. If the p-value is statistically significant (<0.05), end the experiment.
5. If the p-value is not statistically significant draw one more sample for X and Y and retest the hypothesis.
6. Repeat step 5 until you achieve a statistically significant result or until your sample size is 40.

Monte Carlo Distribution



(a) Distribution of p-value for fixed sampling size



(b) Distribution of p-value for continued sampling

Experiment:	Mean sample size	p-value < 0.05
Normal	20.0	0.0494
Optimised	38.1	0.1355

Table 4.2: Results for fixed and varying sample sizes

Bias in Publication

Bias in Publication

- Journals have a tendency to publish papers with only positive results.
- A form of survivorship bias occurs.

Bias in Publication

- Journals have a tendency to publish papers with only positive results.
- A form of survivorship bias occurs.
- Meta-analysis is carried out to account for the largest amount of data.
- But the meta-analysis will only take into account the positive results and so it will be extremely distorted.

Multiple Testing

Multiple Testing

Definition

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses.

Multiple Testing

Definition

Multiple testing refers to any instance that involves the simultaneous testing of several hypotheses.

Example:

- Generate 20 vectors, $X_j's$, each of sample size 100.

$$X_{j,i} \overset{iid}{\sim} N(0,1) \quad \text{with} \quad j = 1, ..., 20$$

- Apply linear regression to each combination of the $X_j's$:

$$X_{j_1} = \beta_0 + \beta_1 X_{j_2} + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0,1)$$

- Where $j_1 \neq j_2$
- H_0: $\beta_1 = 0$

Multiple Testing

$$\binom{20}{2} = 190$$

- 13 out of 190 hypothesis tests were statistically significant.

Multiple Testing

$$\binom{20}{2} = 190$$

- 13 out of 190 hypothesis tests were statistically significant.
- Aware of the possibility of multiple testing.
- We adjust the p-value.

Multiple Testing

$$\binom{20}{2} = 190$$

- 13 out of 190 hypothesis tests were statistically significant.
- Aware of the possibility of multiple testing.
- We adjust the p-value.

Method	Significance Level	False Positives
Original	0.05000	13
Bonferroni	0.00026	1
Harmonic mean	0.00530	1
Order of Magnitude	0.00500	1

Table: Adjusted Significance Level

Multiple testing

- These methods only work if we know that we are testing multiple hypotheses.
- Before we test the hypotheses, we look at our data to see if there are any linear relationships between our variables.

Multiple Testing

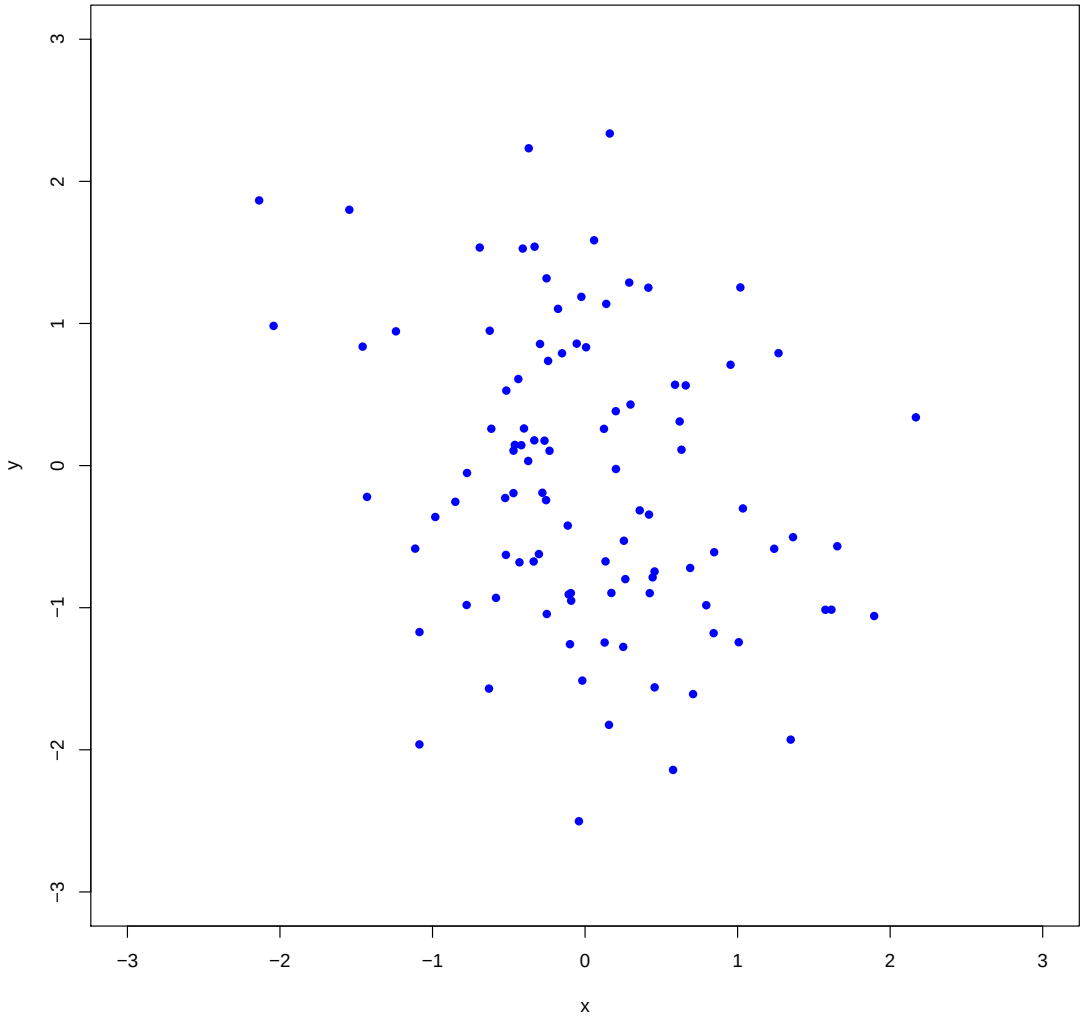


Figure: X_2 and X_{10} plotted against each other

Multiple Testing

- We carry out one hypothesis test.
- Apply linear regression:

$$X_{2,i} = \beta_0 + \beta_1 X_{10,i} + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0,1)$$

- H_0: $\beta_1 = 0$

Multiple Testing

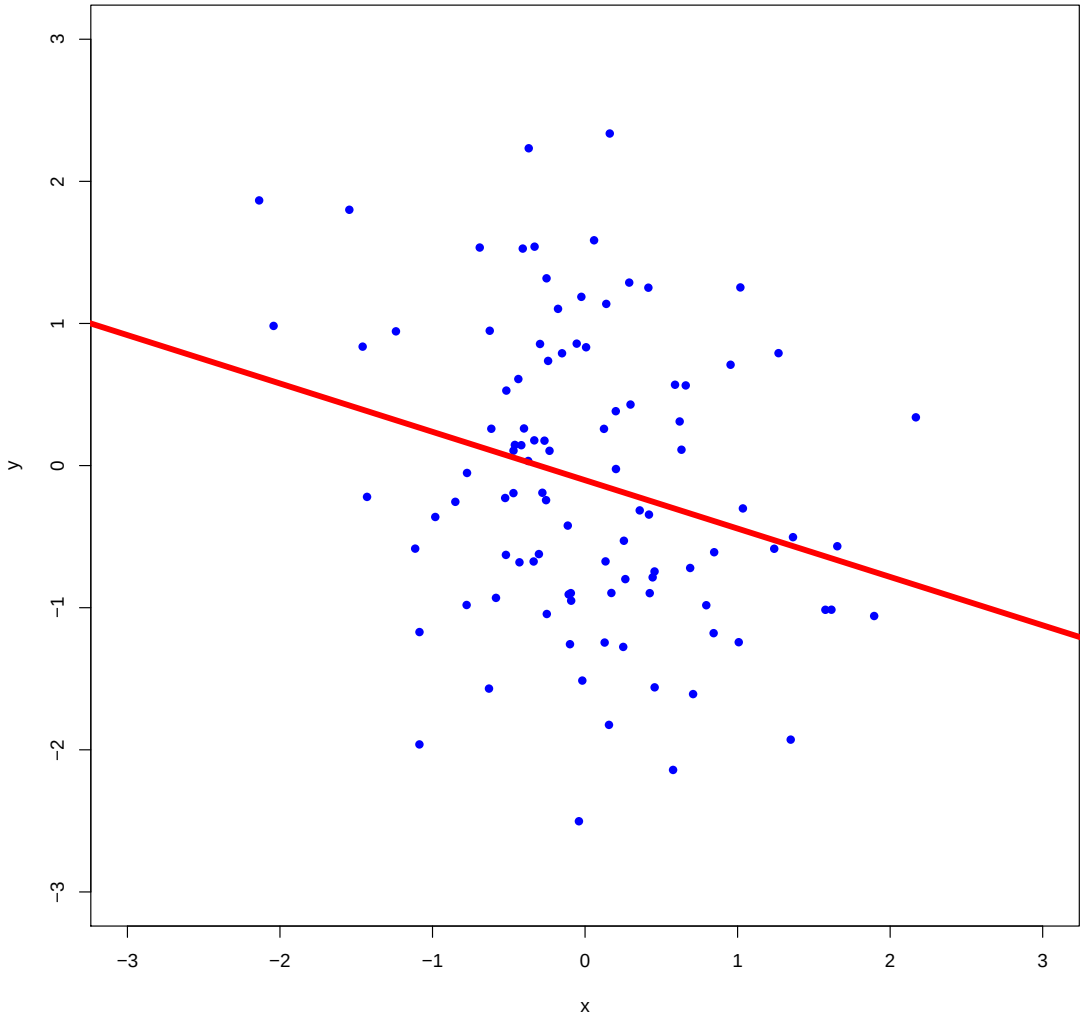


Figure: Model plotted with the data.

Multiple Testing

- We get a statistically significant result (p-value < 0.05).
- By observing our data first, we explicitly carried out only one hypothesis.
- But we implicitly carried out 190 hypothesis testing without adjusting our level of significance.

Curse of Dimensionality

Curse of Dimensionality

Definition

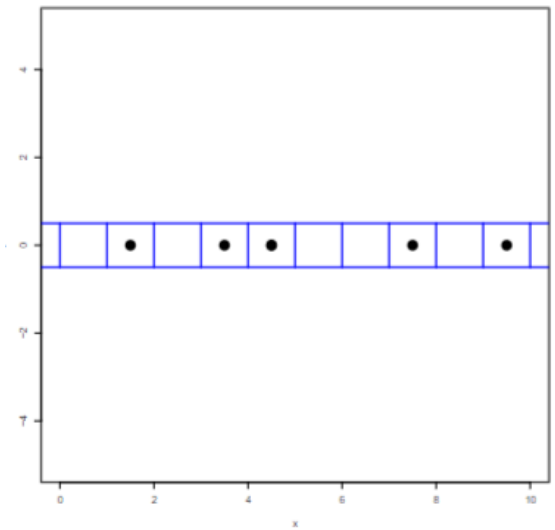
The **Curse of Dimensionality** is to increase features while keeping constant accuracy, we need to increase the amount of data and from the other side to increase the accuracy while keeping the data constant we need to decrease the number of features.

Curse of Dimensionality

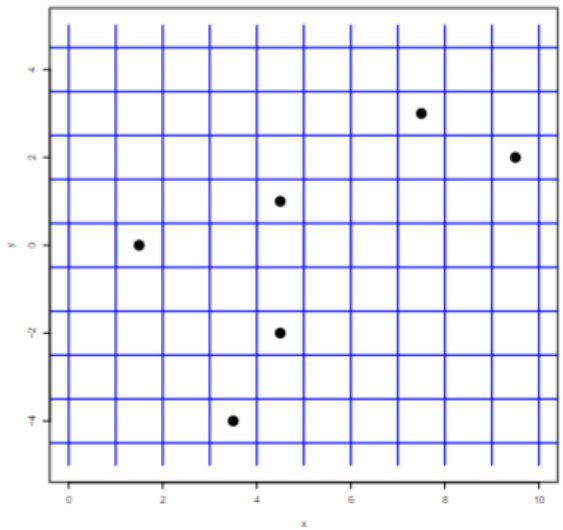
Definition

The **Curse of Dimensionality** is to increase features while keeping constant accuracy, we need to increase the amount of data and from the other side to increase the accuracy while keeping the data constant we need to decrease the number of features.

As the number of features or dimensions grows, the amount of data we need to generalise accuracy grows exponentially



(a) One dimensional with 5 data points



(b) Two dimensional with 5 data points

Feature Elimination

Feature Elimination

Example:
Create data set $S = (X, Y)$

$$X_{1,i} \in \{1, 2, 3, 4\} \qquad X_{2,i} \overset{iid}{\sim} Pois(6)$$

$$X_{3,i} \overset{iid}{\sim} 5 - exp(2) \qquad X_{j,i} \overset{iid}{\sim} N(0, 1) \quad \text{with} \quad j = 4, ..., 23$$

.

Feature Elimination

Example:
Create data set $S = (X, Y)$

$$X_{1,i} \in \{1, 2, 3, 4\} \qquad X_{2,i} \overset{iid}{\sim} Pois(6)$$

$$X_{3,i} \overset{iid}{\sim} 5 - exp(2) \qquad X_{j,i} \overset{iid}{\sim} N(0, 1) \quad \text{with} \quad j = 4, ..., 23$$

. With

$$Y_i = 2\exp(0.15X_{2,i}) + 3.2\log(X_{1,i}) + 4X_{3,i}$$

Feature Elimination

Example:
Create data set $S = (X, Y)$

$$X_{1,i} \in \{1, 2, 3, 4\} \qquad X_{2,i} \overset{iid}{\sim} Pois(6)$$

$$X_{3,i} \overset{iid}{\sim} 5 - exp(2) \qquad X_{j,i} \overset{iid}{\sim} N(0, 1) \quad \text{with} \quad j = 4, ..., 23$$

. With

$$Y_i = 2\exp(0.15X_{2,i}) + 3.2\log(X_{1,i}) + 4X_{3,i}$$

Create models using different feature elimination methods to deal with these 23 features. We can see the percentage of times these methods selected each feature in 10-fold cross validation.

Feature Elimination

Variable	LASSO	Exhaustive	Backward Selection	Random Forests
X_1	100	100	100	100
X_2	100	100	100	100
X_3	100	100	100	100
X_4	50	40	0	0
X_5	10	40	0	70
X_6	40	30	0	0
X_7	0	0	0	10
X_8	10	10	20	30
X_9	90	90	0	10
X_{10}	20	80	10	20
X_{11}	10	0	0	0
X_{12}	60	100	10	10
X_{13}	0	0	70	10
X_{14}	40	100	0	10
X_{15}	0	0	0	50
X_{16}	70	70	0	50
X_{17}	0	10	20	60
X_{18}	0	0	100	20
X_{19}	60	90	70	0
X_{20}	0	0	0	0
X_{21}	0	0	10	0
X_{22}	80	100	0	0
X_{23}	0	10	90	50

Consequences

If we cannot trust the results of our scientific research then many aspects of our society may be compromised:

Consequences

If we cannot trust the results of our scientific research then many aspects of our society may be compromised:

- Public health
  - Medical treatment
  - Nutritional information.
- Industry
  - Finance
  - Agriculture
- Government Policy
  - Healthcare
  - Education

Consequences

If we cannot trust the results of our scientific research then many aspects of our society may be compromised:

- Public health
  - Medical treatment
  - Nutritional information.
- Industry
  - Finance
  - Agriculture
- Government Policy
  - Healthcare
  - Education
- **Loss of trust in Science**

Conclusion

Thank you for listening


Ronald Fisher.
Statistical methods and scientific induction.
Journal of the Royal Statistical Society. Series B (Methodological), 17(1):69–78, 1955.


Gerd Gigerenzer.
Mindless statistics.
The Journal of Socio-Economics, 33(5):587–606, 2004.
Statistical Significance.


J. Neyman.
"inductive behavior" as a basic concept of philosophy of science.
Revue de l'Institut international de statistique, 25(1/3):7, 1957.


Nassim Nicholas Taleb.
A short note on p-value hacking, 2018.