# Introduction

Maarten Speekenbrink

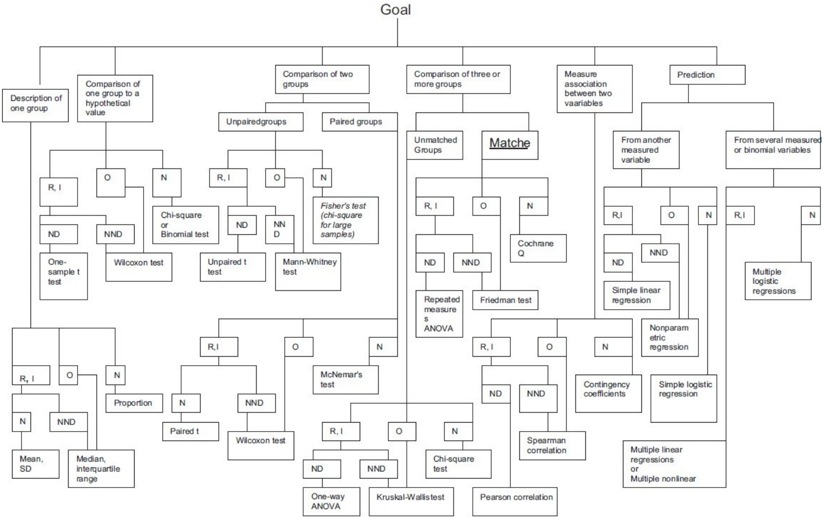
Experimental Psychology
University College London

Statistics: lecture 1

# Outline

1. Course overview
2. Distributions
3. Statistical models
4. Estimation
5. Prediction

# The ‘cookbook’ approach




R, I = Ratio and Interval data  O= Ordinal data  N = Nominal data

N = Normal distribution  NND = Non normal distribution

# The model comparison approach

- One general framework
  - Statistical models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{p-1} X_{p-1,i} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma)$$

  - Compare prediction error between different models
- Focus on building models of data
  - Flexibility: build model to your needs
  - But: need to think!
- Most common statistical techniques fit in this framework
  - t-test
  - Multiple linear regression
  - ANOVA
  - ANCOVA
  - Repeated measures ANOVA
  - Mixed effects models
  - Logistic regression

# Course objectives

- Refresh/introduce basic statistical concepts
- Solid understanding of the General Linear Model (GLM)
- Learn to build models to investigate your data
- Learn to formulate hypothesis tests as comparisons between statistical models
- Critically evaluate model assumptions

# Structure of the course

| | |
|---|---|
| Monday lecture: | introducing topics |
| Homework: | exercise sheet |
| Wednesday/Thursday lab: | SPSS (and possibly R) |
| Online screencasts: | review exercise sheet |

Course material on
`http://moodle.ucl.ac.uk/course/view.php?id=11131`
Moodle course name: `PSYCGR01: Statistics`
Enrolment key: `STATS2017`
Guest access: `STATISTICS17`

# Assessment

The assessment of the course consists of three computer-based (Moodle) exams taken throughout the course. The two best scored exams (out of three) will each count towards 50% of your final mark.

Exam dates:

| | | |
|---|---|---|
| Tuesday 31 October 2017 | 9:30-10:30 | Exam 1 |
| Tuesday 28 November 2017 | 9:30-10:30 | Exam 2 |
| Tuesday 12 December 2017 | 9:30-10:30 | Exam 3 |

# Literature

Main textbook:

- Judd, C.M., McClelland, G.H. & Ryan, C.S. (2009). *Data Analysis: A Model Comparison Approach To Regression, ANOVA, and Beyond* (3rd ed.). Routledge (you will get by with the 2nd edition)
- Articles and other material made available via Moodle

More general sources that some may find helpful:

- Field, A. *Discovering Statistics Using SPSS*
- Howell, D. C. *Statistical methods for psychology*
- Abelson, R. P. *Statistics as principled argument*
- Foster, J.J., Barkus, E. & Yavorsky, C. *Understanding and using advanced statistics*

Also have a look at the links on the wiki on Moodle...

# Today's objectives

- Refresh and review some basic concepts:
  - Sampling & distributions
  - Estimation
  - Prediction

# What is statistics?

**Wikipedia**

"Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data." (Wikipedia)

# What is statistics?

**Wikipedia**

"Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data." (Wikipedia)

**One I like better:**

"Statistics is the science of learning from limited and noisy data"

# What is statistics?

### Wikipedia

"Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data." (Wikipedia)

### One I like better:

"Statistics is the science of learning from limited and noisy data"

What can I do with statistics?

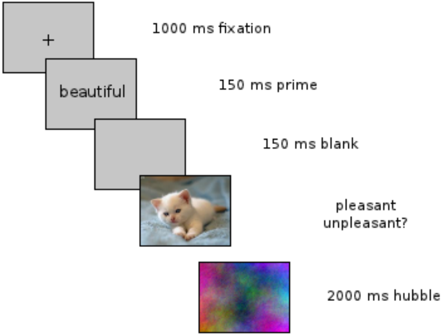
- Evaluate the empirical support for theoretical claims
- Make informed decisions

What does it not do?

- Tell you The Truth

# Feeling the future (Bem, 2011)

Affective priming

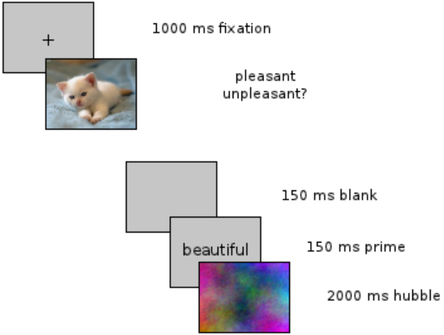



Primes: pleasant, unpleasant
Pictures: pleasant, unpleasant
Faster response for congruent primes

Bem, D.J. (2011) Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425

# Feeling the future (Bem, 2011)

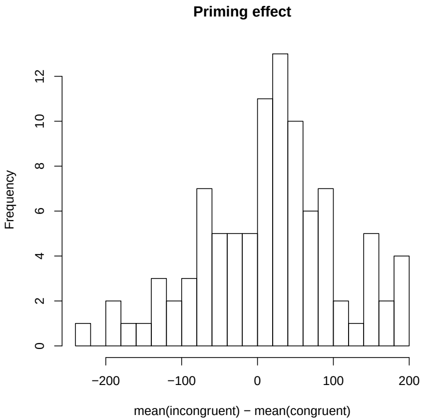
Retroactive priming




Primes: pleasant, unpleasant
Pictures: pleasant, unpleasant
Faster response for congruent primes ?

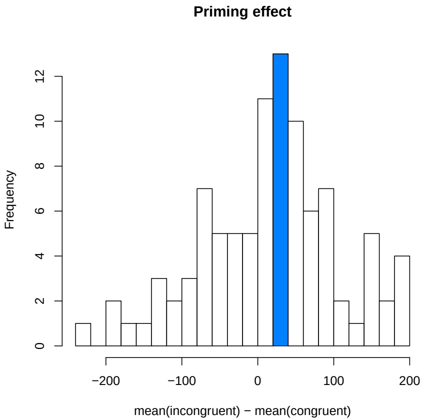
Bem, D.J. (2011) Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425

# Feeling the future (simulated data)




$$\begin{aligned} Y_i &= \text{priming effect for participant } i \\ &= \text{mean(RT incongr.)} - \text{mean(RT congr.)} \end{aligned}$$

# Measures of location



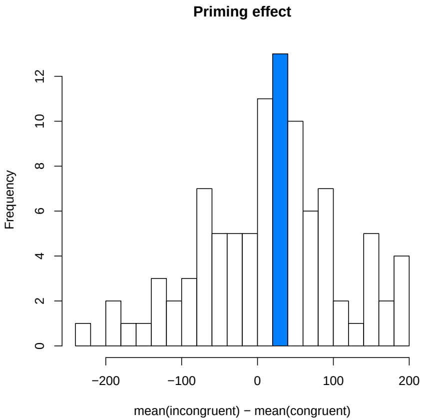

$Y_i$ = priming effect for participant $i$

= mean(RT incongr.) – mean(RT congr.)

Mode (most frequent value)
For this histogram: the bar between 20-40

(doesn't actually exists for this data, as no values are the same, but does exist for histogram)

# Measures of location



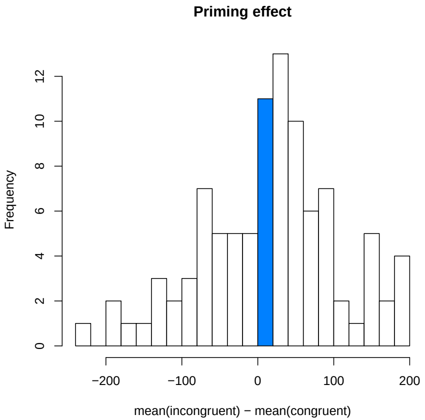

$Y_i$ = priming effect for participant $i$

= mean(RT incongr.) – mean(RT congr.)

Median (value such that 50% of values are lower)

$$\text{Me}_Y = 23.16$$

# Measures of location




$Y_i$ = priming effect for participant $i$

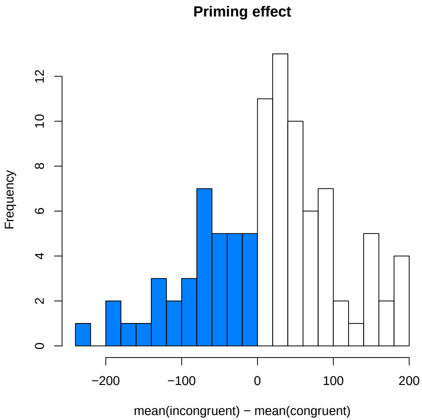
= mean(RT incongr.) – mean(RT congr.)

Mean (average value)

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = 15.58$$

($n$ is the total number of observations)

# Measures of spread

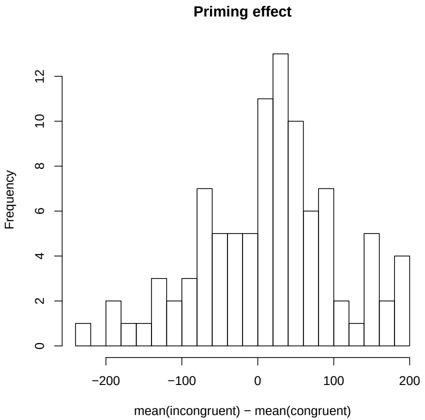



$Y_i$ = priming effect for participant $i$

= mean(RT incongr.) – mean(RT congr.)

Priming effect not always positive (negative for 36% of participants)

# Measures of spread



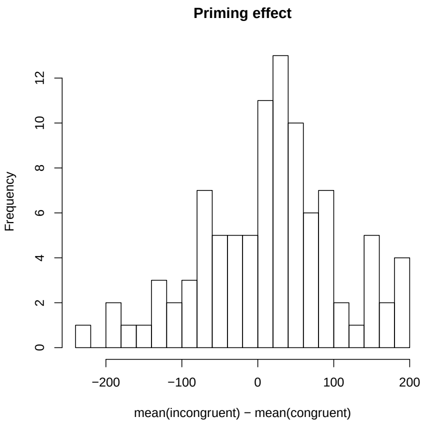

$Y_i =$ priming effect for participant $i$

$= \text{mean(RT incongr.)} - \text{mean(RT congr.)}$

Variance:

$$S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}$$
$$= 8152.04$$

# Measures of spread




$Y_i =$ priming effect for participant $i$

$=$ mean(RT incongr.) – mean(RT congr.)

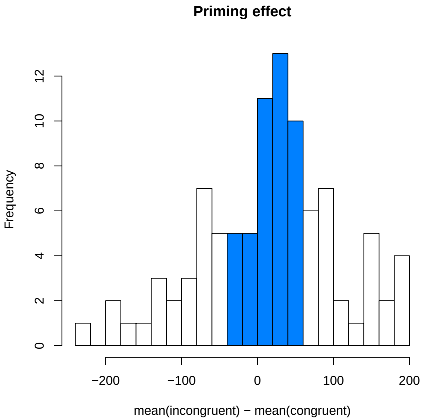
Standard deviation:

$$S_Y = \sqrt{S_Y^2}$$

$$= \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n}}$$

$$= 90.28$$

# Measures of spread



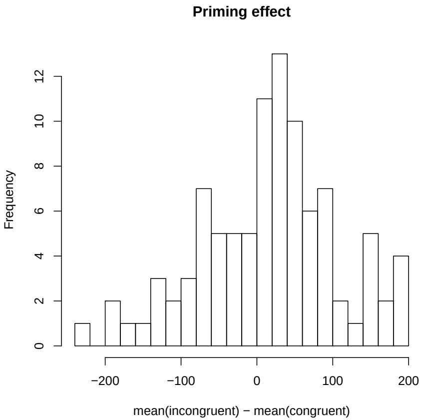

$Y_i$ = priming effect for participant $i$

= mean(RT incongr.) – mean(RT congr.)

Interquartile range (range such that 25% of the data is below, and 25% of the data is above the range)

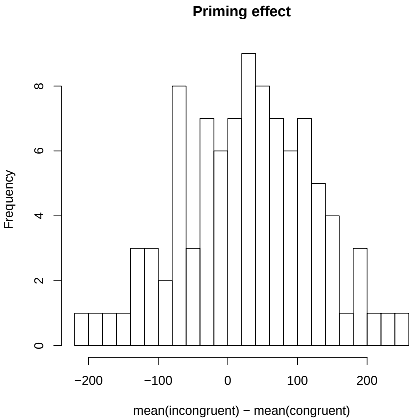
Interquartile range: $[-43.12; 63.24]$

# Sample and population distribution




Observed data is only one of many possible datasets

# Sample and population distribution



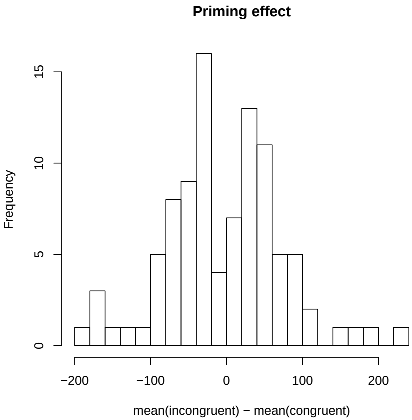

Observed data is only one of many possible datasets

If the experiment was repeated with other people, or at a different time, the data could look have looked different. . .

$$\overline{Y} = 27.05$$

$$S_Y = 96.02$$

# Sample and population distribution

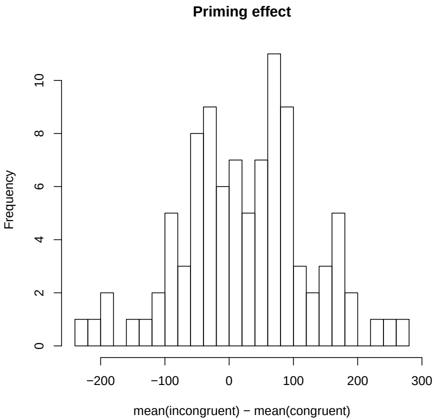



Observed data is only one of many possible datasets

If the experiment was repeated with other people, or at a different time, the data could look have looked different. . .

$$\overline{Y} = -2.88$$
$$S_Y = 76.70$$

# Sample and population distribution



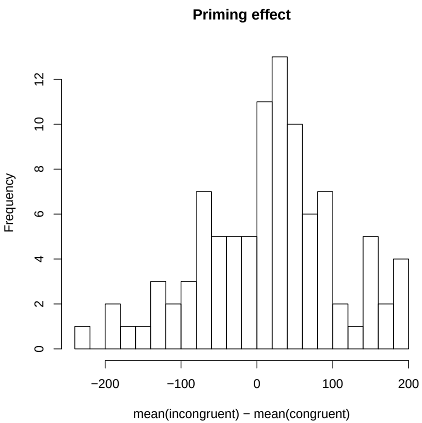

Observed data is only one of many possible datasets

If the experiment was repeated with other people, or at a different time, the data could look have looked different. . .

$$\overline{Y} = 26.00$$

$$S_Y = 99.36$$

# Sample and population distribution




Observed data is only one of many possible datasets

If the experiment was repeated with other people, or at a different time, the data could look have looked different. . .

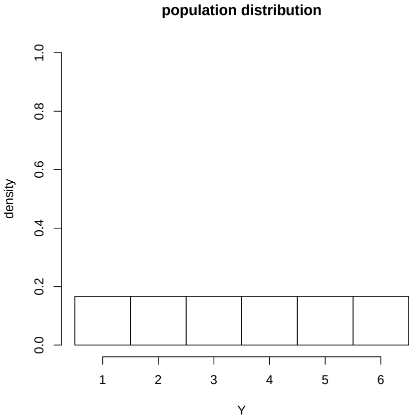
Both the sample mean and variance vary from sample to sample!

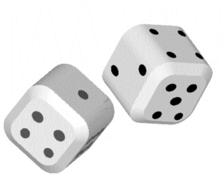
# Sample and population distribution

- Data is a sample from the "population", a (hypothetical) construct defined as the collection of all possible observations (e.g., from all people at all times).
- We'd like to say something about the population (e.g., its mean), not about the sample.
- Because the data is a random sample from the population, we can use the data to infer properties of the population. This is *statistical inference*.

# Throwing dice (slight detour)

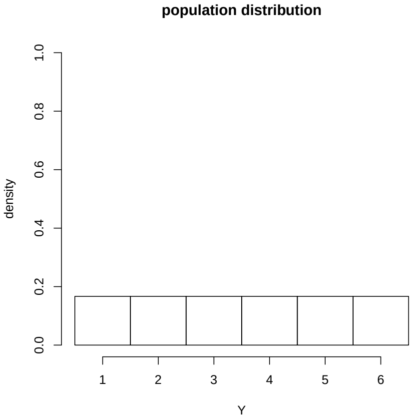
Population distribution
(single die)



# Throwing dice (slight detour)

Population distribution (single die)




Mean:

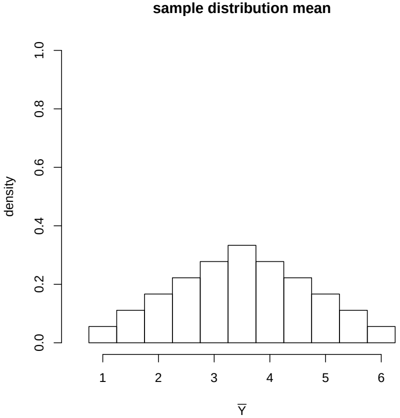
$$\mu = 3.5$$

Variance:

$$\sigma^2 \approx 2.92$$

# Throwing dice (slight detour)

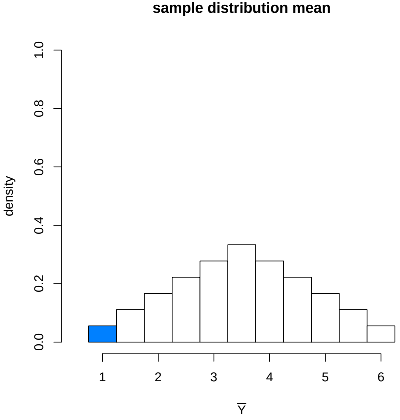
Sampling distribution mean (two dice)




| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
| 2 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
| 3 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |
| 4 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
| 5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 |
| 6 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 |

# Throwing dice (slight detour)

Sampling distribution mean (two dice)

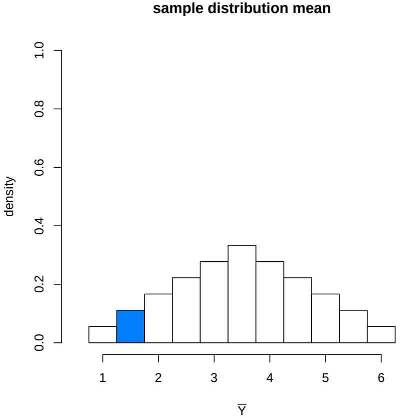



| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
| 2 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
| 3 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |
| 4 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
| 5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 |
| 6 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 |

$$p(\overline{Y} = 1) = \frac{1}{36}$$

# Throwing dice (slight detour)

Sampling distribution mean
(two dice)

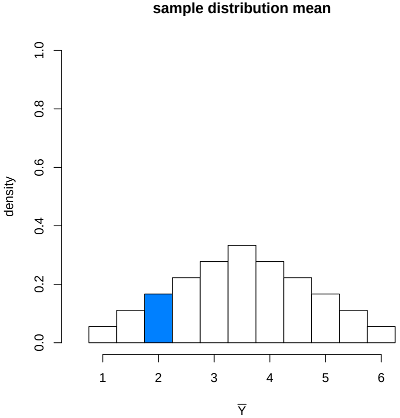



| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
| 2 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
| 3 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |
| 4 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
| 5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 |
| 6 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 |

$$p(\overline{Y} = 1.5) = \frac{2}{36}$$

# Throwing dice (slight detour)

Sampling distribution mean (two dice)

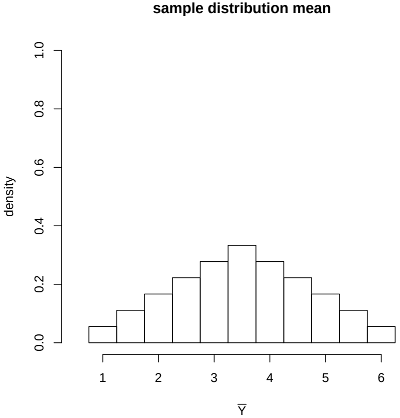



| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
| 2 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
| 3 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |
| 4 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
| 5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 |
| 6 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 |

$$p(\overline{Y} = 2) = \frac{3}{36}$$

# Throwing dice (slight detour)

Sampling distribution mean (two dice)




| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
| 2 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
| 3 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |
| 4 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
| 5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 |
| 6 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 |

Mean:

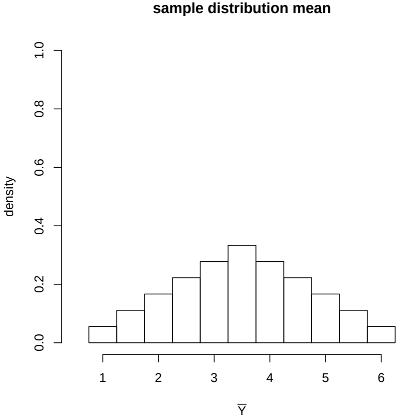
$$\mu = 3.5$$

Variance:

$$\sigma^2 \approx 1.46$$

# Throwing dice (slight detour)

Sampling distribution mean (two dice)

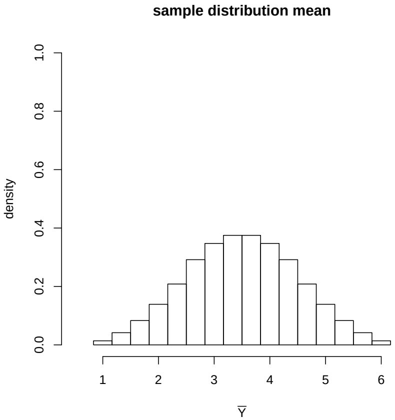



| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 |
| 2 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 |
| 3 | 2 | 2.5 | 3 | 3.5 | 4 | 4.5 |
| 4 | 2.5 | 3 | 3.5 | 4 | 4.5 | 5 |
| 5 | 3 | 3.5 | 4 | 4.5 | 5 | 5.5 |
| 6 | 3.5 | 4 | 4.5 | 5 | 5.5 | 6 |

This theoretical distribution of the sample mean is also called the sampling distribution of the mean.

# Throwing dice (slight detour)

Sampling distribution mean (three dice)




Mean:

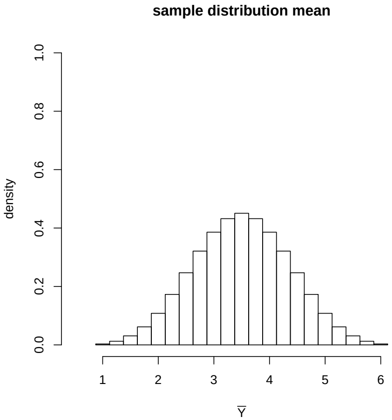
$$\mu = 3.5$$

Variance:

$$\sigma^2 \approx 0.97$$

# Throwing dice (slight detour)

Sampling distribution mean (four dice)




Mean:

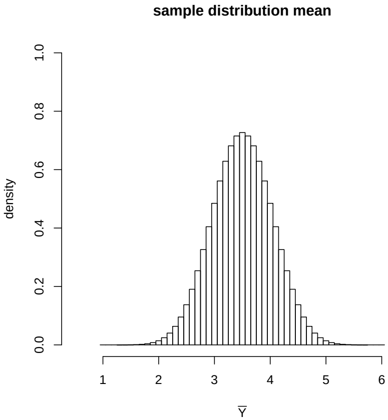
$$\mu = 3.5$$

Variance:

$$\sigma^2 \approx 0.73$$

# Throwing dice (slight detour)

Sampling distribution mean (ten dice)




Mean:

$$\mu = 3.5$$

Variance:

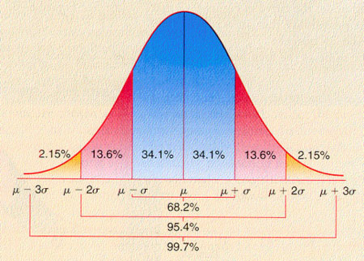
$$\sigma^2 \approx 0.29$$

# Normal distribution

Population is commonly assumed to follow a normal distribution. The Normal probability density function is defined as

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}}$$

It has two parameters, the mean $\mu$ and standard deviation $\sigma$.




- For the Normal distribution: mean = median = mode
- The assumption of a Normal distribution can be justified by the Central Limit Theorem.

# Central Limit Theorem

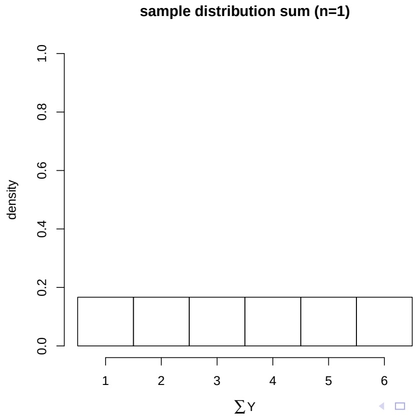
## Central limit theorem

Under mild conditions, the distribution of a sum of $n$ independent random variables $Y_i$ converges to a Normal distribution as $n$ goes to infinity

# Central Limit Theorem

## Central limit theorem

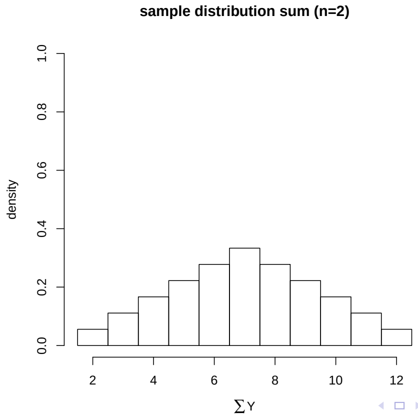
Under mild conditions, the distribution of a sum of $n$ independent random variables $Y_i$ converges to a Normal distribution as $n$ goes to infinity



# Central Limit Theorem

## Central limit theorem

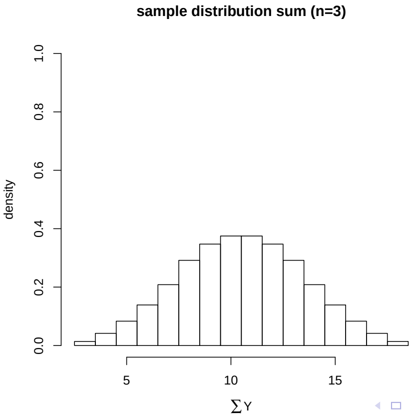
Under mild conditions, the distribution of a sum of $n$ independent random variables $Y_i$ converges to a Normal distribution as $n$ goes to infinity



# Central Limit Theorem

## Central limit theorem

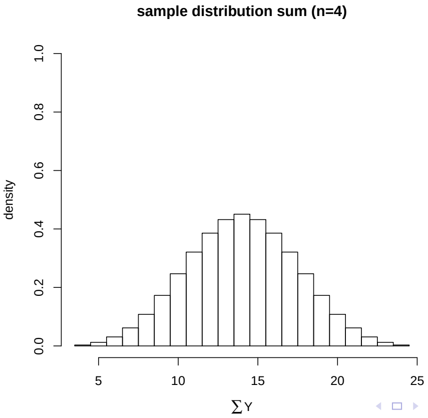
Under mild conditions, the distribution of a sum of $n$ independent random variables $Y_i$ converges to a Normal distribution as $n$ goes to infinity



# Central Limit Theorem

## Central limit theorem

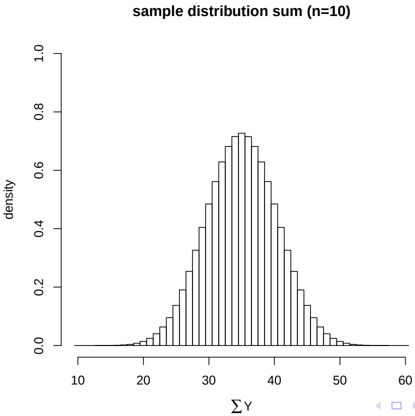
Under mild conditions, the distribution of a sum of $n$ independent random variables $Y_i$ converges to a Normal distribution as $n$ goes to infinity



# Central Limit Theorem

## Central limit theorem

Under mild conditions, the distribution of a sum of $n$ independent random variables $Y_i$ converges to a Normal distribution as $n$ goes to infinity



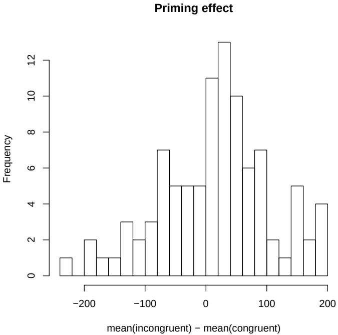
# Statistical models

- Assumption of a Normal distribution is a statistical model of the data

$$Y_i \sim N(\mu, \sigma)$$

  (read as "$Y_i$ is sampled from a Normal distribution with mean $\mu$ and standard deviation $\sigma$"). The model has two parameters: $\mu$ and $\sigma$
- The model is (quite likely) an idealization
  - balance between complexity and descriptive validity
  - "all models are wrong, but some are useful" (G.E.P. Box)
- A model is useful insofar as it
  - helps to understand a phenomenon
  - provides good predictions for future data

# A statistical model of retroactive priming

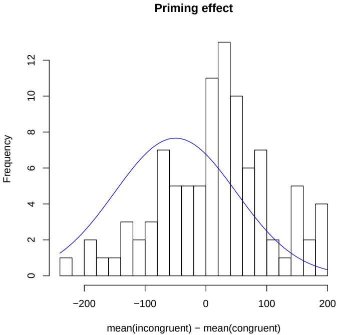



- We have data $Y_i$, $i = 1, \ldots, N$, and a model

$$Y_i \sim N(\mu, \sigma)$$

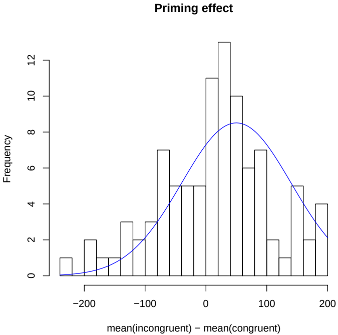
  which describes the population the sample was drawn from.
- We don't know the values of $\mu$ and $\sigma$, but we can estimate them from the sample.
- For the Normal distribution, mean = median = mode. Which of these should we use to estimate $\mu$?

# A statistical model of retroactive priming




- We have data $Y_i$, $i = 1, \ldots, N$, and a model
$$Y_i \sim N(\mu, \sigma)$$
which describes the population the sample was drawn from.
- We don't know the values of $\mu$ and $\sigma$, but we can estimate them from the sample.
- For the Normal distribution, mean = median = mode. Which of these should we use to estimate $\mu$?

# A statistical model of retroactive priming

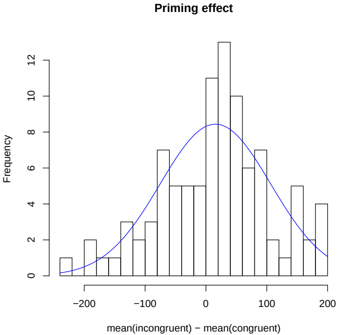



- We have data $Y_i$, $i = 1, \ldots, N$, and a model

$$Y_i \sim N(\mu, \sigma)$$

  which describes the population the sample was drawn from.
- We don't know the values of $\mu$ and $\sigma$, but we can estimate them from the sample.
- For the Normal distribution, mean = median = mode. Which of these should we use to estimate $\mu$?

# A statistical model of retroactive priming

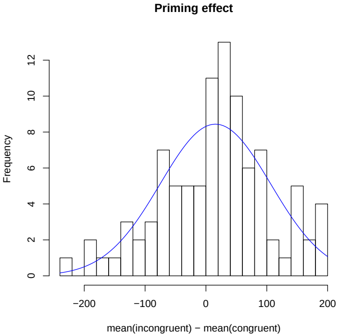



- We have data $Y_i$, $i = 1, \ldots, N$, and a model

$$Y_i \sim N(\mu, \sigma)$$

  which describes the population the sample was drawn from.
- We don't know the values of $\mu$ and $\sigma$, but we can estimate them from the sample.
- For the Normal distribution, mean = median = mode. Which of these should we use to estimate $\mu$?

# A statistical model of retroactive priming




- We have data $Y_i$, $i = 1, \ldots, N$, and a model

$$Y_i \sim N(\mu, \sigma)$$

  which describes the population the sample was drawn from.
- We don't know the values of $\mu$ and $\sigma$, but we can estimate them from the sample.
- For the Normal distribution, mean = median = mode. Which of these should we use to estimate $\mu$?

# Sampling distribution of the mean

The sampling distribution of the mean $\overline{Y}$ of $n$ independent samples from a Normal distribution with mean $\mu$ and standard deviation $\sigma$ is a Normal distribution with mean $\mu_{\overline{Y}} = \mu$ and standard deviation $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$:

$$Y_i \sim N(\mu, \sigma) \qquad \overline{Y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- On average, the sample mean is equal to the population mean.
- Large deviations between the sample and population mean become less likely with larger sample size $n$ (the “law of large numbers”).

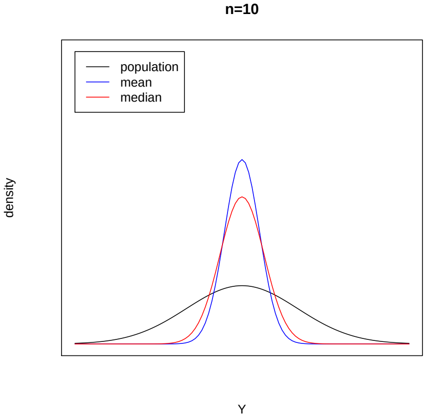
# Sampling distribution of the median

The sampling distribution of the median $\text{Me}_Y$ of $n$ independent samples from a Normal distribution with mean (and median) $\mu$ and standard deviation $\sigma$ is (approximately, for large $n$) a Normal distribution with mean $\mu$ and standard deviation $\sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{n}}$:

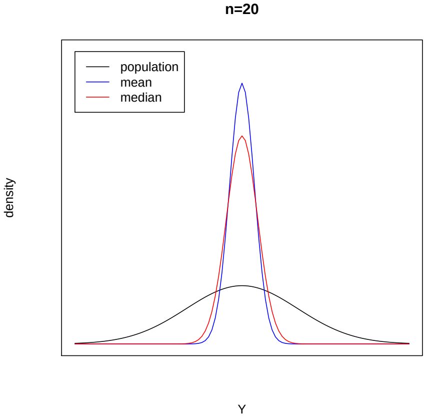
$$Y_i \sim N(\mu, \sigma) \qquad \text{Me}_Y \sim N\left(\mu, \sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

- On average, the sample median is equal to the population median (and mean).
- Large deviations between the sample and population median become less likely with larger sample size $n$ (law of large numbers).

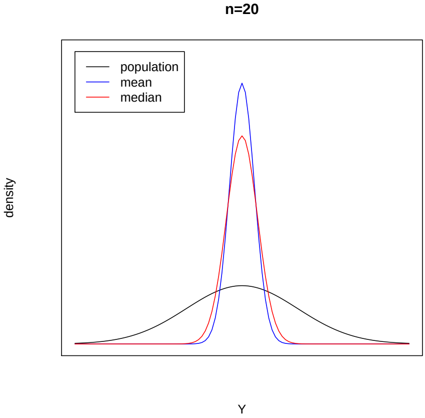
# Comparing the distribution of the mean and median



# Comparing the distribution of the mean and median



# Comparing the distribution of the mean and median




With increasing $n$, both become better estimators. But the variance of the median is $\frac{\pi}{2} \approx 1.57$ times larger than the variance of the mean

# How good is an estimator?

Estimates have a sampling distribution. Judge the quality of an estimator from this distribution:

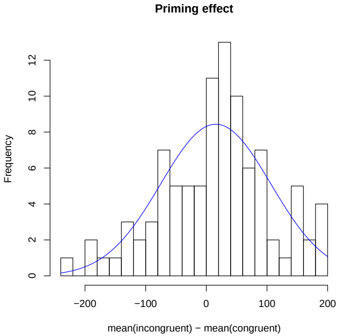
- Unbiasedness: mean of sampling distribution is the true value.
- Consistency: variance decreases with increasing sample size.
- Efficiency: smallest possible variance out of all unbiased estimators.

| | Unbiased | Consistent | Efficient |
|---|---|---|---|
| Sample mean | ✓ | ✓ | ✓ |
| Sample median | ✓ | ✓ | ✗ |
| Sample variance | ✗ | ✓ | ✗ |

An unbiased and efficient estimator of the population variance is

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{n-1}$$

# A statistical model of retroactive priming

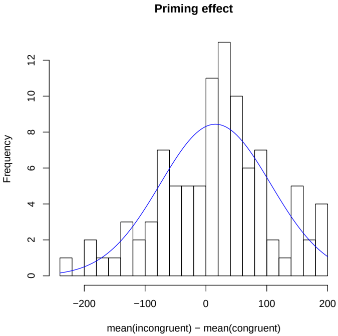



- We have data $Y_i$ and model

$$Y_i \sim N(\mu, \sigma)$$

  We do not know $\mu$ and $\sigma$
- We can estimate the parameters from the data
  - $\hat{\mu} = \overline{Y} = 15.58$, $\hat{\sigma} = 90.28$
- We can use the model to make various predictions:
  - On average, the priming effect will be 15.58
  - On average, 43% of priming effects will be negative
- Note: these predictions are based on an estimated model!

# A statistical model of retroactive priming



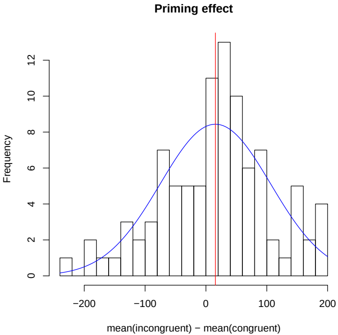

- We have data $Y_i$ and model

$$Y_i \sim N(\mu, \sigma)$$

  We do not know $\mu$ and $\sigma$
- We can estimate the parameters from the data
  - $\hat{\mu} = \overline{Y} = 15.58$, $\hat{\sigma} = 90.28$

- We can use the model to make various predictions:
  - On average, the priming effect will be 15.58
  - On average, 43% of priming effects will be negative
- Note: these predictions are based on an estimated model!

# A statistical model of retroactive priming



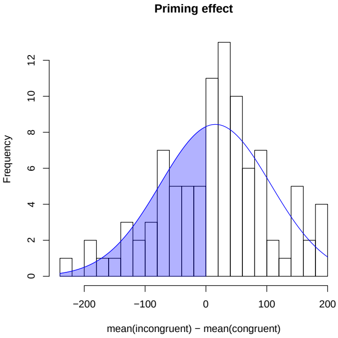

- We have data $Y_i$ and model

$$Y_i \sim N(\mu, \sigma)$$

  We do not know $\mu$ and $\sigma$
- We can estimate the parameters from the data
  - $\hat{\mu} = \overline{Y} = 15.58$, $\hat{\sigma} = 90.28$

- We can use the model to make various predictions:
  - On average, the priming effect will be 15.58
  - On average, 43% of priming effects will be negative
- Note: these predictions are based on an estimated model!

# A statistical model of retroactive priming



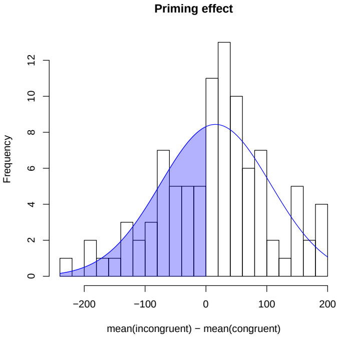

- We have data $Y_i$ and model

$$Y_i \sim N(\mu, \sigma)$$

  We do not know $\mu$ and $\sigma$
- We can estimate the parameters from the data
  - $\hat{\mu} = \overline{Y} = 15.58$, $\hat{\sigma} = 90.28$

- We can use the model to make various predictions:
  - On average, the priming effect will be 15.58
  - On average, 43% of priming effects will be negative
- Note: these predictions are based on an estimated model!

# A statistical model of retroactive priming



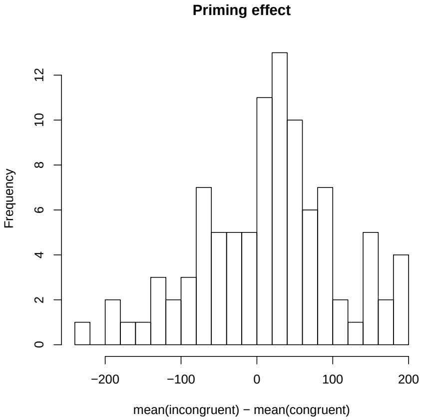

- We have data $Y_i$ and model

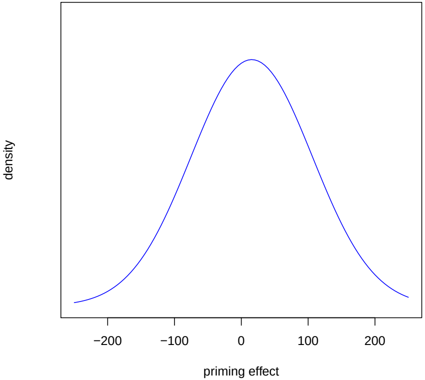
$$Y_i \sim N(\mu, \sigma)$$

  We do not know $\mu$ and $\sigma$
- We can estimate the parameters from the data
  - $\hat{\mu} = \overline{Y} = 15.58$, $\hat{\sigma} = 90.28$

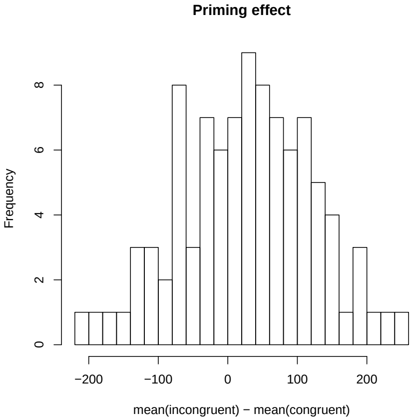
- We can use the model to make various predictions:
  - On average, the priming effect will be 15.58
  - On average, 43% of priming effects will be negative
- Note: these predictions are based on an estimated model!

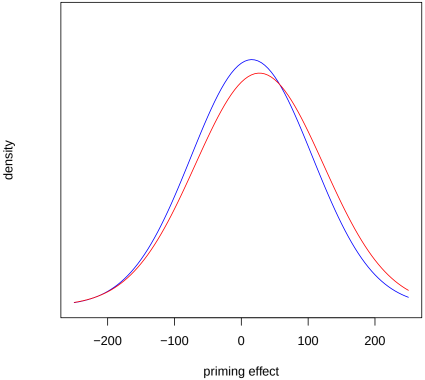
# Different samples, different estimated models. . .



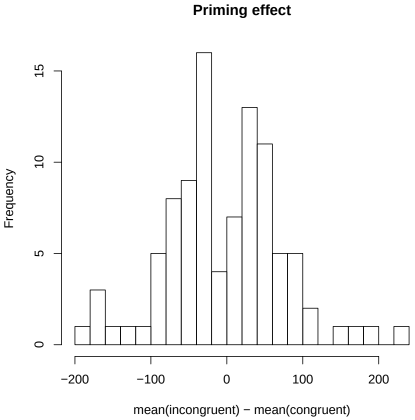



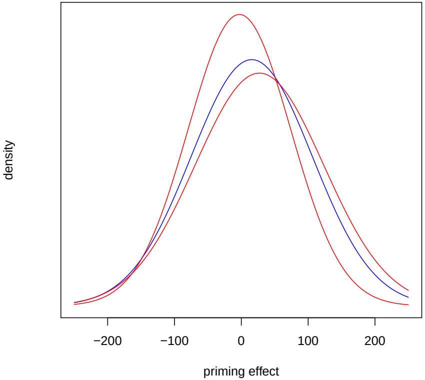
# Different samples, different estimated models. . .



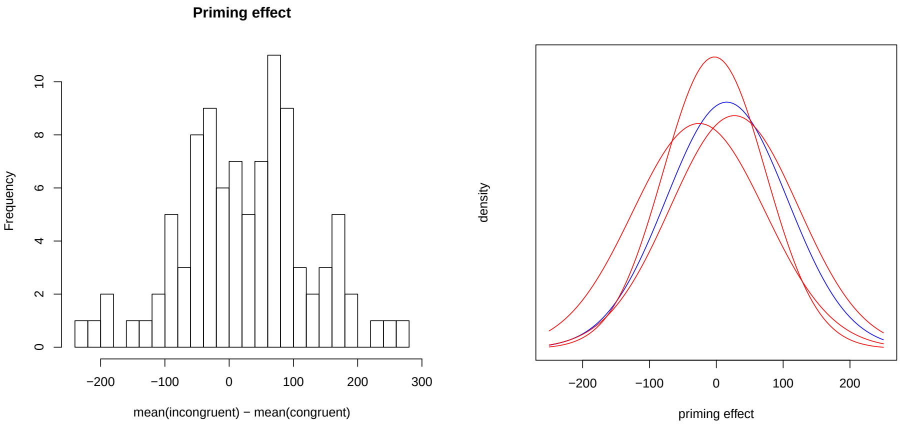



# Different samples, different estimated models...



# Different samples, different estimated models. . .



# Prediction

We have data $Y_i$. A prediction is denoted as $\hat{Y}_i$. You can always say

$$Y_i = \hat{Y}_i + e_i$$

where $e_i = Y_i - \hat{Y}_i$ is the prediction error

If we don't know anything specific about data point $Y_i$, we can use the (estimated) population mean as our prediction, i.e. $\hat{Y}_i = \overline{Y}$:

| $Y_i$ | $\hat{Y}_i$ | $e_i$ |
|---|---|---|
| -238.98 | 15.58 | -254.56 |
| 21.16 | 15.58 | 5.58 |
| 140.18 | 15.58 | 124.60 |
| 7.71 | 15.58 | -7.87 |
| | ⋮ | |

# Prediction error

| $Y_i$ | $\hat{Y}_i$ | $e_i$ |
|---|---|---|
| -238.98 | 15.58 | -254.56 |
| 21.16 | 15.58 | 5.58 |
| 140.18 | 15.58 | 124.60 |
| 7.71 | 15.58 | -7.87 |
| $\vdots$ | $\vdots$ | $\vdots$ |

A good model has small prediction errors. Different error measures

- Sum of errors: $\sum_{i=1}^{n} e_i$
- Count of errors: $\sum_{i=1}^{n} I(e_i \neq 0)$
- Sum of absolute errors: $\sum_{i=1}^{n} |e_i|$
- Sum of squared errors: $\sum_{i=1}^{n} e_i^2$

# Prediction error

| $Y_i$ | $\hat{Y}_i$ | $e_i$ |
|---:|---:|---:|
| -238.98 | 15.58 | -254.56 |
| 21.16 | 15.58 | 5.58 |
| 140.18 | 15.58 | 124.60 |
| 7.71 | 15.58 | -7.87 |
| ⋮ | ⋮ | ⋮ |
| | $\sum =$ | 0 |

A good model has small prediction errors. Different error measures

- Sum of errors: $\sum_{i=1}^{n} e_i$
- Count of errors: $\sum_{i=1}^{n} I(e_i \neq 0)$
- Sum of absolute errors: $\sum_{i=1}^{n} |e_i|$
- Sum of squared errors: $\sum_{i=1}^{n} e_i^2$

# Prediction error

| $Y_i$ | $\hat{Y}_i$ | $e_i$ | $I(e_i \neq 0)$ |
|---|---|---|---|
| -238.98 | 15.58 | -254.56 | 1 |
| 21.16 | 15.58 | 5.58 | 1 |
| 140.18 | 15.58 | 124.60 | 1 |
| 7.71 | 15.58 | -7.87 | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $\sum =$ | 0 | 96 |

A good model has small prediction errors. Different error measures

- Sum of errors: $\sum_{i=1}^{n} e_i$
- Count of errors: $\sum_{i=1}^{n} I(e_i \neq 0)$
- Sum of absolute errors: $\sum_{i=1}^{n} |e_i|$
- Sum of squared errors: $\sum_{i=1}^{n} e_i^2$

# Prediction error

| $Y_i$ | $\hat{Y}_i$ | $e_i$ | $I(e_i \neq 0)$ | $\lvert e_i \rvert$ |
|---|---|---|---|---|
| -238.98 | 15.58 | -254.56 | 1 | 254.56 |
| 21.16 | 15.58 | 5.58 | 1 | 5.58 |
| 140.18 | 15.58 | 124.60 | 1 | 124.60 |
| 7.71 | 15.58 | -7.87 | 1 | 7.87 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $\sum =$ | 0 | 96 | 6653.19 |

A good model has small prediction errors. Different error measures

- Sum of errors: $\sum_{i=1}^{n} e_i$
- Count of errors: $\sum_{i=1}^{n} I(e_i \neq 0)$
- Sum of absolute errors: $\sum_{i=1}^{n} \lvert e_i \rvert$
- Sum of squared errors: $\sum_{i=1}^{n} e_i^2$

# Prediction error

| $Y_i$ | $\hat{Y}_i$ | $e_i$ | $I(e_i \neq 0)$ | $\|e_i\|$ | $e_i^2$ |
|---:|---:|---:|---:|---:|---:|
| -238.98 | 15.58 | -254.56 | 1 | 254.56 | 64800.79 |
| 21.16 | 15.58 | 5.58 | 1 | 5.58 | 31.14 |
| 140.18 | 15.58 | 124.60 | 1 | 124.60 | 15525.16 |
| 7.71 | 15.58 | -7.87 | 1 | 7.87 | 61.94 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $\sum =$ | 0 | 96 | 6653.19 | 782599.6 |

A good model has small prediction errors. Different error measures

- Sum of errors: $\sum_{i=1}^{n} e_i$
- Count of errors: $\sum_{i=1}^{n} I(e_i \neq 0)$
- Sum of absolute errors: $\sum_{i=1}^{n} |e_i|$
- Sum of squared errors: $\sum_{i=1}^{n} e_i^2$

# Prediction error in the sample and population

The model is estimated from the data and we can compute the prediction errors for this data. We can estimate the model to minimize the prediction errors. This leads to different estimates, depending on the error measure used:

- Count of errors: mode
- Sum of absolute errors: median
- Sum of squared errors: mean

Usually use Sum of Squared Errors (SSE):

- Many small errors are better than one large error
- Related to variance
- Easier to work with mathematically

# Prediction error in the population

- While the estimated model minimizes the SSE in the sample, it is unlikely to minimize the SSE in the population
  - SSE in the sample minimized when $\hat{Y}_i = \overline{Y}$
  - SSE in the population is minimized when $\hat{Y}_i = \mu$
- Sampling distribution of $\overline{Y}$ guarantees that $\overline{Y} = \mu$ *on average*, but for any given sample, $\overline{Y}$ is unlikely to equal $\mu$ exactly
- So while $\hat{Y}_i = \overline{Y}$ is the best prediction for the current sample from the population, it is unlikely to be the best prediction for the whole population
- But if we don't have any further information about $\mu$, it is the best we can do... Next lecture we'll look at what to do when you do have a particular hypothesis about $\mu$...

# Summary of key points

- Observed data is a sample from a population. The population is characterized by the population distribution
  - Distributions are characterized by parameters such as the mean, median, mode, variance, standard deviation, *etc*.
- The population is a hypothetical construct that constitutes a statistical model for the data. Statistical models can be useful to derive sampling distributions, make predictions for future data, understand where your observations come from, *etc*.
- The parameters of a statistical model are often unknown, but can be estimated from sample data. Good estimators are unbiased, consistent, and efficient
- A good model provides good predictions. To estimate a model, we can minimize the prediction error
  - Usually Sum of Squared Error (SSE), but others possible
  - Minimum prediction error in sample does not guarantee minimum prediction error for the population!

# Further reading

This lecture:

- Judd, McClelland & Ryan chapters 1 - 3

For next lecture:

- Judd, McClelland & Ryan chapter 4