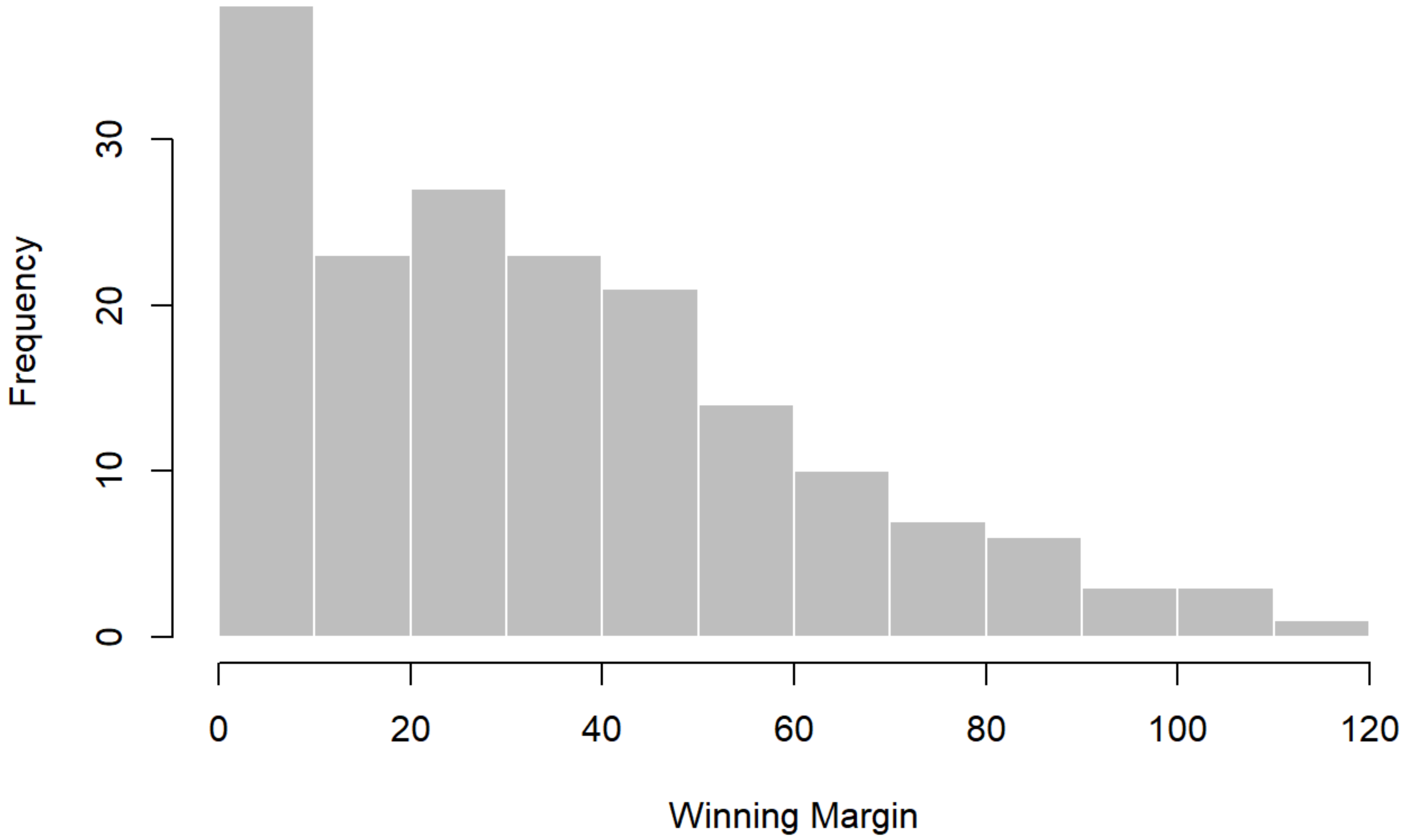
Stats for AI

Rishabh Kumar

Topics to cover

- **Summarising Statistics**
- **Probability Theory**
- **Bayes Theorem**
- Distributions
- Samples
- Hypothesis Testing
- Comparing means
- Linear Regression
- What is AI
- Simple ML Overview
- Deep Learning Overview
- Reinforcement Learning Overview

Data



Mean

the observation	its symbol	the observed value
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points
winning margin, game 3	X_3	56 points
winning margin, game 4	X_4	8 points
winning margin, game 5	X_5	32 points

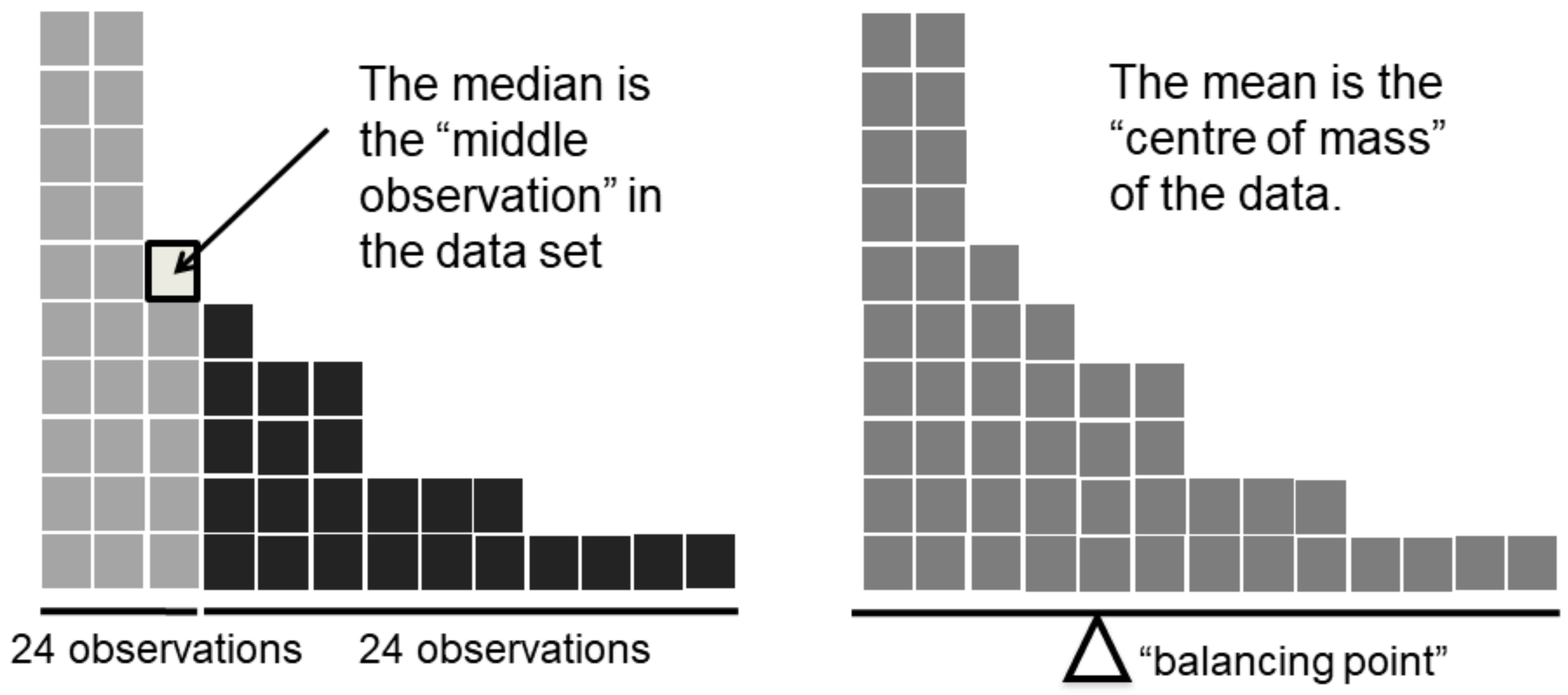
$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{N-1} + X_N}{N}$$

$$\sum_{i=1}^{5} X_i$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Median

$$8, 31, \mathbf{32}, 56, 56$$



Bankers using mean and median

Senior Commonwealth Bank executives have travelled the world in the past couple of weeks with a presentation showing how Australian house prices, and the key price to income ratios, compare favourably with similar countries. “Housing affordability has actually been going sideways for the last five to six years,” said Craig James, the chief economist of the bank’s trading arm, CommSec.

CBA has waged its war against what it believes are housing doomsayers with graphs, numbers and international comparisons. In its presentation, the bank rejects arguments that Australia’s housing is relatively expensive compared to incomes. It says Australia’s house price to household income ratio of 5.6 in the major cities, and 4.3 nationwide, is comparable to many other developed nations. It says San Francisco and New York have ratios of 7, Auckland’s is 6.7, and Vancouver comes in at 9.3.

Many analysts say that has led the bank to use misleading figures and comparisons. If you go to page four of CBA’s presentation and read the source information at the bottom of the graph and table, you would notice there is an additional source on the international comparison – Demographia. However, if the Commonwealth Bank had also used Demographia’s analysis of Australia’s house price to income ratio, it would have come up with a figure closer to 9 rather than 5.6 or 4.3

[An] obvious problem with the Commonwealth Bank’s domestic price to income figures is they compare average incomes with median house prices (unlike the Demographia figures that compare median incomes to median prices). The median is the mid-point, effectively cutting out the highs and lows, and that means the average is generally higher when it comes to incomes and asset prices, because it includes the earnings of Australia’s wealthiest people. To put it another way: the Commonwealth Bank’s figures count Ralph Norris’ multi-million dollar pay packet on the income side, but not his (no doubt) very expensive house in the property price figures, thus understating the house price to income ratio for middle-income Australians.

Trimmed Mean

Cut the extremes

Group A = -100, 3,5,7,20

Group B = - 15,0,1,4,7,12

Highly sensitive to one or two extreme values, and is thus not considered to be a ***robust*** measure.

Mode

The mode of a sample is very simple: it is the value that occurs most frequently

Measures of Variability

Range

The ***range*** of a variable is the biggest value minus the smallest value.

Interquartile range

The ***interquartile range*** (IQR) is the difference between the 25th quantile and the 75th quantile. Probably you already know what a ***quantile***

It is like the range, but instead of calculating the difference between the biggest and smallest value it focuses on 25th and 7th quantiles

Mean absolute deviation

A different approach is to select a meaningful reference point (usually the mean or the median) and then report the "typical" deviations from that reference point

the observation	its symbol	the observed value
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points
winning margin, game 3	X_3	56 points
winning margin, game 4	X_4	8 points
winning margin, game 5	X_5	32 points

$$\frac{19.4 + 5.6 + 19.4 + 28.6 + 4.6}{5} = 15.52$$

$$(X) = \frac{1}{N} \sum_{i=1}^{N} |X_i - \bar{X}|$$

Variance (σ^2) and standard deviation (σ)

Squared mean deviation

i	Xi Value	X_i - mean	$(X_i$ - mean)^2
1	56	19.4	376.36
2	31	-5.6	31.36
3	56	19.4	376.36
4	8	-28.6	817.96
5	32	-4.6	21.16

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Z = A + B

Var(Z) = Var(A) + Var(B)

Standard Scores or z scores

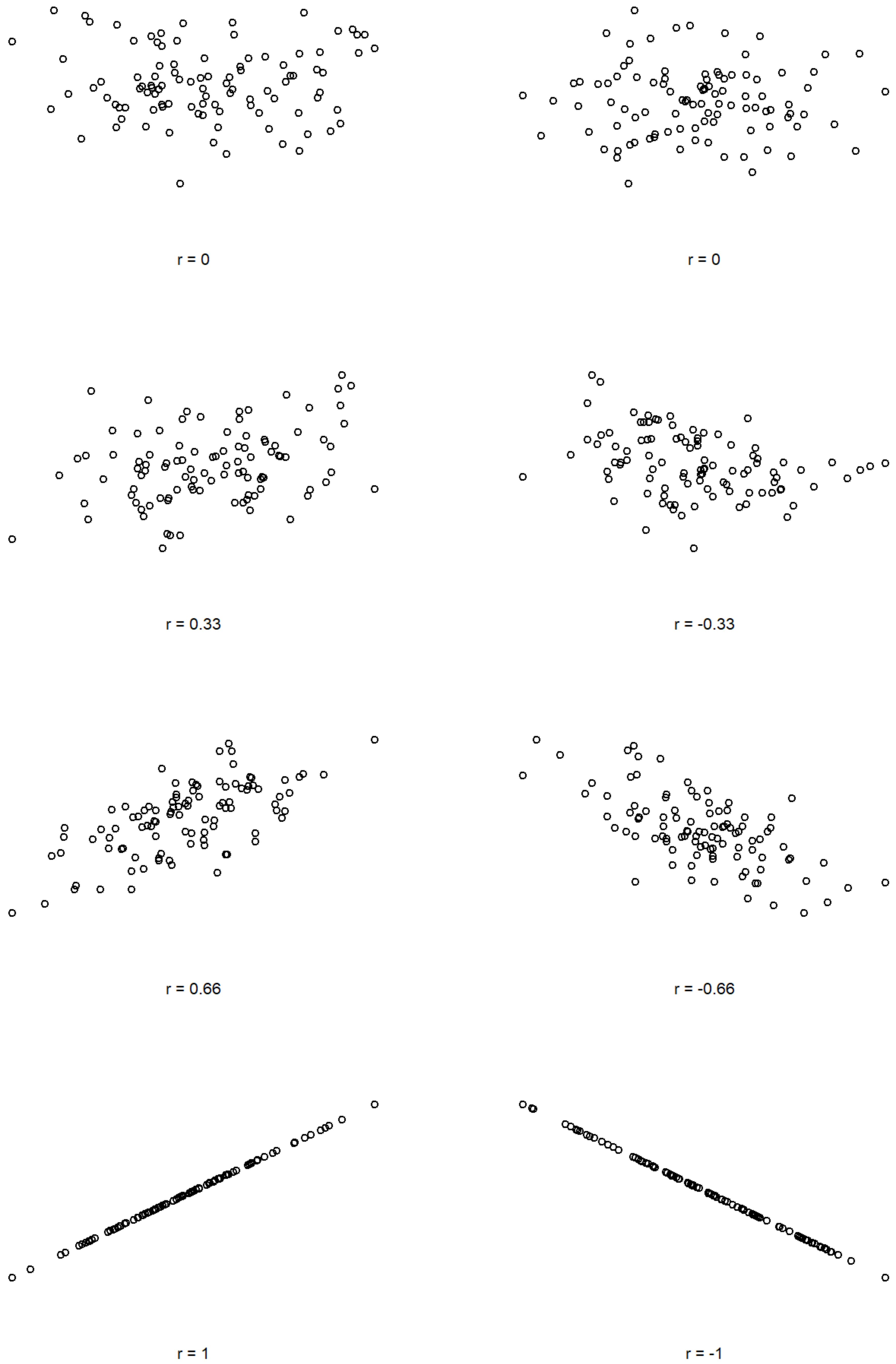
A different approach is to convert my grumpiness score into a ***standard score***, also referred to as a *z score*
z score. The standard score is defined as the number of standard deviations above the mean that my grumpiness score lies.

$$standard\ score = \frac{(raw\ score - mean)}{\sigma}$$

$$z_i = \frac{(X_i - \bar{X})}{\hat{\sigma}}$$

$$z_i = \frac{(35 - 17)}{5} = 3.6$$

Correlations



Probability Theory

Inferential statistics provides the tools that we need to answer these sorts of questions, and since these kinds of questions lie at the heart of the scientific enterprise.

However, the theory of statistical inference is built on top of ***probability theory***

All statistical theory is built upon probability but most staticians dont agree what probability is

What is probability

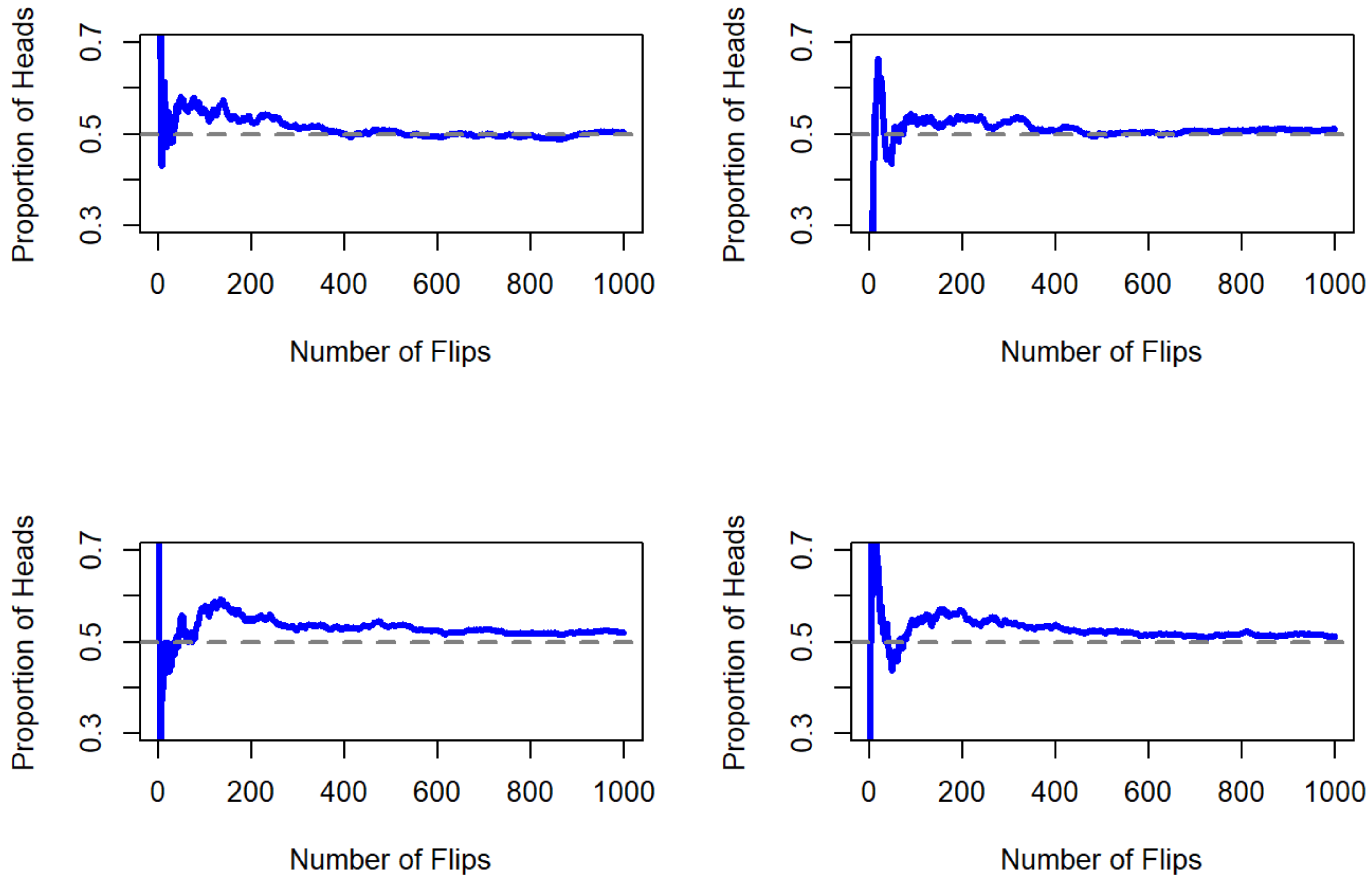
- They're robot teams, so I can make them play over and over again, and if I did that, *Arduino Arsenal* would win 8 out of every 10 games on average. **[Odds]**

- For any given game, I would only agree that betting on this game is only "fair" if a $1 bet on *C Milan* gives a $5 payoff (i.e. I get my $1 back plus a $4 reward for being correct), as would a $4 bet on *Arduino Arsenal* (i.e., my $4 bet plus a $1 reward). **[Expectations]**

- My subjective "belief" or "confidence" in an *Arduino Arsenal* victory is four times as strong as my belief in a *C Milan* victory **[Beliefs]**

Frequentist

- The more dominant view in statistics, is referred to as the ***frequentist view***, and it defines probability as a ***long-run frequency***.



Pros

- Firstly, it is objective: the probability of an event is *necessarily* grounded in the world. The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe.

- Secondly, it is unambiguous: any two people watching the same sequence of events unfold, trying to calculate the probability of an event

Cons

- Firstly, infinite sequences don't exist in the physical world.

- The frequentist definition has a narrow scope. Frequentist probability genuinely *forbids* us from making probability statements about a single event.

- "There is a category of days for which I predict a 60% chance of rain; if we look only across those days for which I make this prediction, then on 60% of those days it will actually rain"

Bayesian

- The ***Bayesian view*** of probability is often called the subjectivist view, and it is a minority view among statisticians, but one that has been steadily gaining traction for the last several decades.

- The most common way of thinking about subjective probability is to define the probability of an event as the ***degree of belief*** that an intelligent and rational agent assigns to that truth of that event.

- The main **advantage** is that it allows you to assign probabilities to any event you want to. You don't need to be limited to those events that are repeatable.

- The main **disadvantage** (to many people) is that we can't be purely objective – specifying a probability requires us to specify an entity that has the relevant degree of belief. This entity might be a human, an alien, a robot, or even a statistician, but there has to be an intelligent agent out there that believes in things.

- To many people this is uncomfortable: it seems to make probability arbitrary.

Bayesian

- From a Bayesian perspective, statistical inference is all about *belief revision*.
- I start out with a set of candidate hypotheses H_s about the world.
- I don't know which of these hypotheses is true, but do I have some beliefs about which hypotheses are plausible and which are not.
- When I observe the data I have to revise those beliefs