Announcements

- HW1 due tonight on Gradescope!
- Lab 1 write-up due tonight as well
- Wednesday will be our delayed in-class lab
  - Working with datasets and plotting
- Read rest of Appendix A.1 for Friday

Warm Up

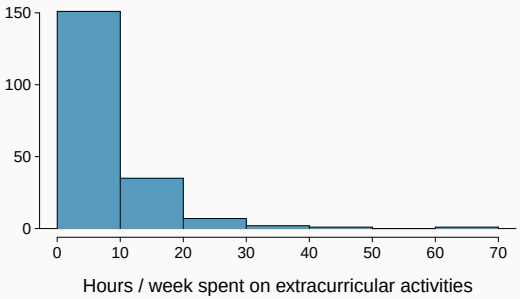
How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?

A) Unimodal and right skewed

B) Unimodal and left skewed

C) Bimodal and symmetric

D) Multimodal and left skewed



Warm Up

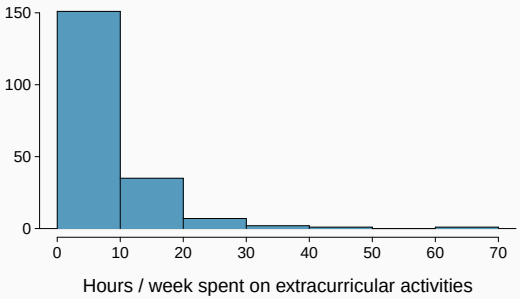
How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?

A) Unimodal and right skewed

B) Unimodal and left skewed

C) Bimodal and symmetric

D) Multimodal and left skewed



Variance

Variance is roughly the average squared deviation from the mean.

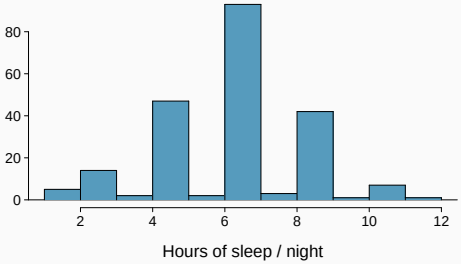
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.

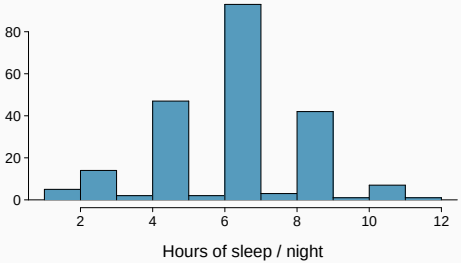


Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11 \ hours^2$$

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

Standard deviation

The ***standard deviation*** is the square root of the variance, and has the same units as the data.

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

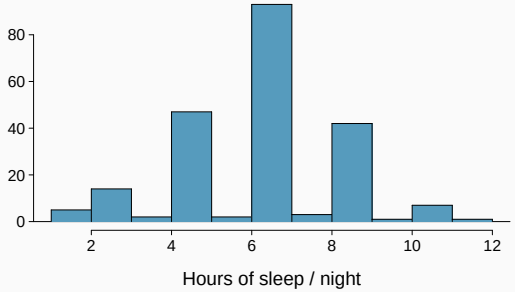
Standard deviation

The ***standard deviation*** is the square root of the variance, and has the same units as the data.

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

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$



Standard deviation

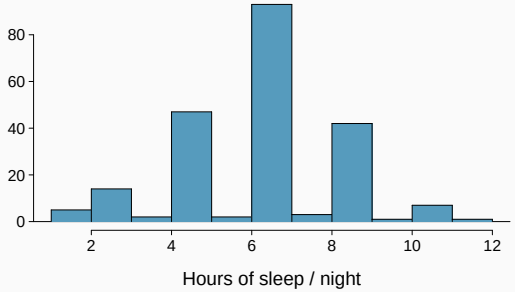
The ***standard deviation*** is the square root of the variance, and has the same units as the data.

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

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$

- We can see that all of the data are within 3 standard deviations of the mean.



Median

- The ***median*** is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50^{th} ***percentile***.

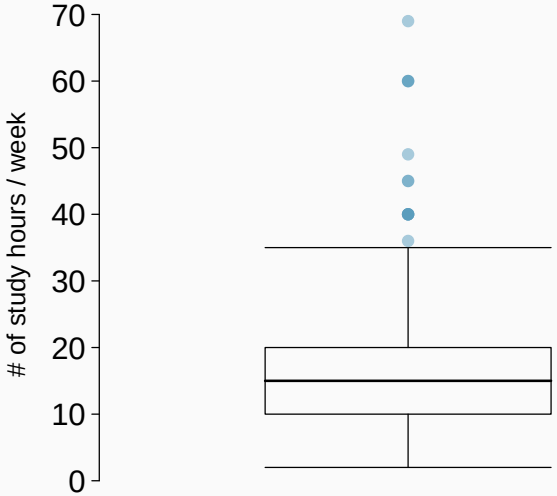
Q1, Q3, and IQR

- The 25^{th} percentile is also called the first quartile, ***Q1***.
- The 50^{th} percentile is also called the median.
- The 75^{th} percentile is also called the third quartile, ***Q3***.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the ***interquartile range***, or the ***IQR***.

$$IQR = Q3 - Q1$$

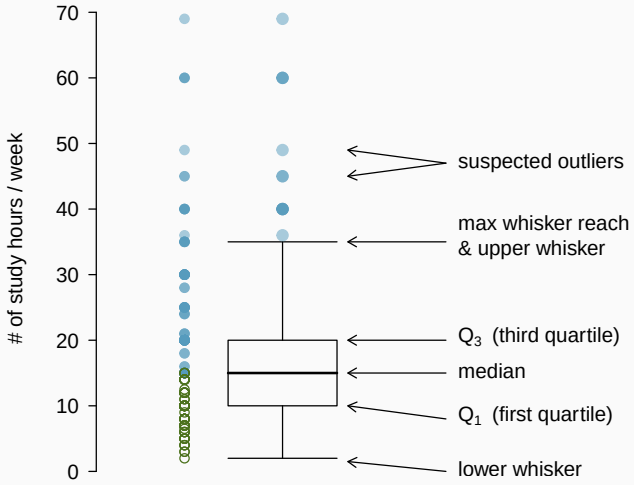
Box plot

The box in a ***box plot*** represents the middle 50% of the data, and the thick line in the box is the median.



Box plot

The box in a ***box plot*** represents the middle 50% of the data, and the thick line in the box is the median.



Whiskers and outliers

- ***Whiskers*** of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

Whiskers and outliers

- ***Whiskers*** of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

Whiskers and outliers

- ***Whiskers*** of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- A potential ***outlier*** is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Why is it important to look for outliers?

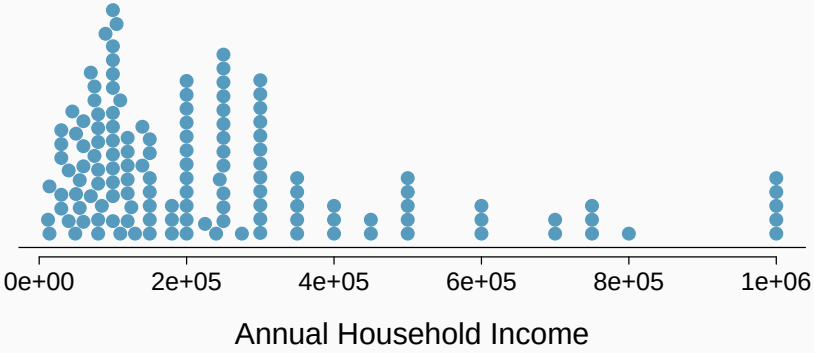
Outliers (cont.)

Why is it important to look for outliers?

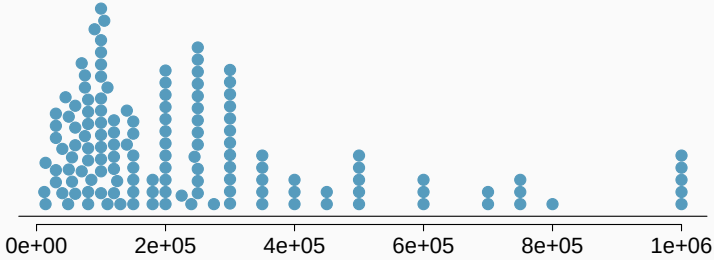
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Extreme observations

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with $10 million? What if the smallest value was replaced with $10 million?



Robust statistics



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original data	190K	200K	245K	226K
move largest to $10 million	190K	200K	309K	853K
move smallest to $10 million	200K	200K	316K	854K

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

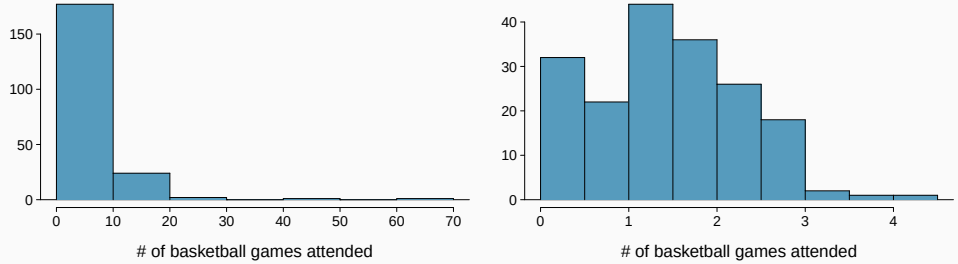
Extremely skewed data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the ***log transformation***.

Extremely skewed data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the ***log transformation***.

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.



Pros and cons of transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70	50	25	$\cdots$
log(# of games)	4.25	3.91	3.22	$\cdots$

- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

Considering categorical data

Contingency tables

A table that summarizes data for two categorical variables is called a ***contingency table***.

Contingency tables

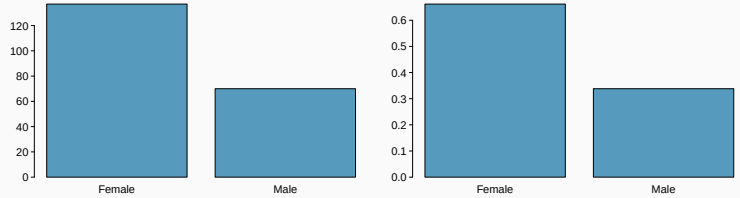
A table that summarizes data for two categorical variables is called a ***contingency table***.

The contingency table below shows the distribution of students' genders and whether or not they are looking for a spouse while in college.

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

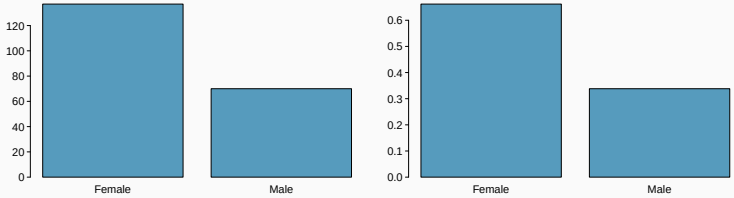
Bar plots

A ***bar plot*** is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a ***relative frequency bar plot***.



Bar plots

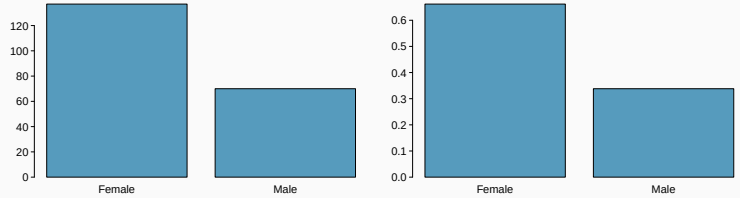
A ***bar plot*** is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a ***relative frequency bar plot***.



How are bar plots different than histograms?

Bar plots

A ***bar plot*** is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a ***relative frequency bar plot***.



How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order.

Choosing the appropriate proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

Choosing the appropriate proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

Choosing the appropriate proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

- % Females looking for a spouse: $51/137 \approx 0.37$

Choosing the appropriate proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

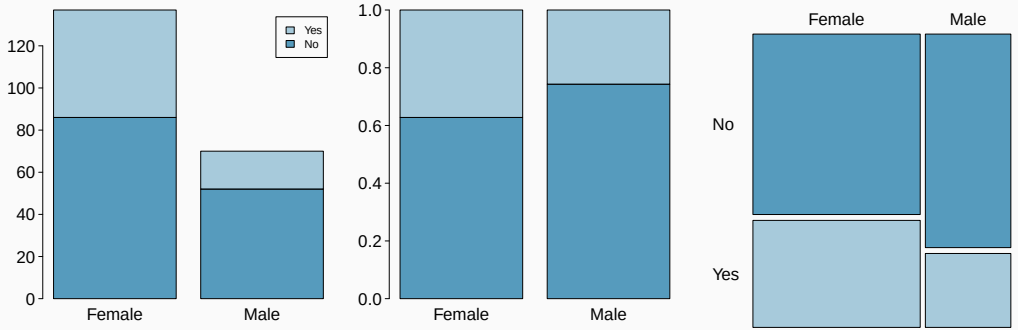
		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

- % Females looking for a spouse: $51/137 \approx 0.37$
- % Males looking for a spouse: $18/70 \approx 0.26$

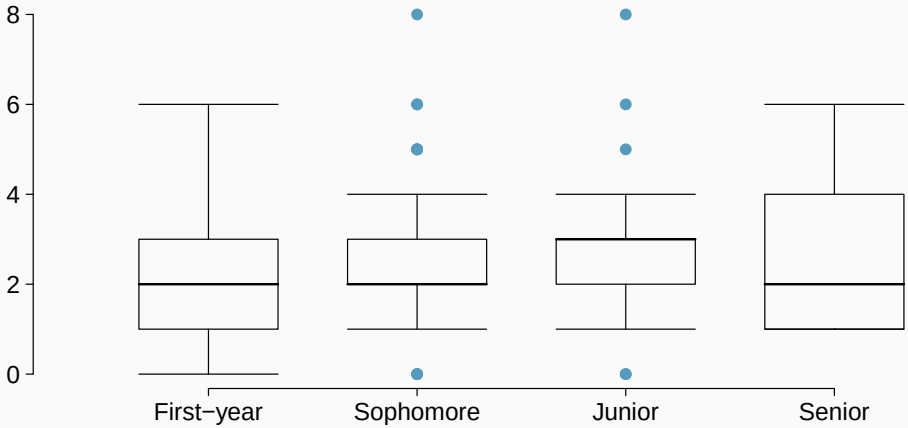
Segmented bar and mosaic plots

What are the differences between the three visualizations shown below?



Side-by-side box plots

Does there appear to be a relationship between class year and number of clubs students are in?



Defining probability

Random processes

- A ***random process*** is a situation in which we know what outcomes could happen, but we don't know which particular outcome will happen.
- Examples: coin tosses, die rolls, iTunes shuffle, whether the stock market goes up or down tomorrow, etc.
- It can be helpful to model a process as random even if it is not truly random.

MP3 Players › Stories › iTunes: Just how random is random?

iTunes: Just how random is random?

By David Braue on 08 March 2007

- Introduction
- Say You, Say What?
- A role for labels?
- The new random

Think that song has appeared in your playlists just a few too many times? David Braue puts the randomness of Apple's song shuffling to the test -- and finds some surprising results.

Quick -- think of a number between one and 20. Now think of another one, and another, and another.

Starting to repeat yourself? No surprise: in practice, many series of random numbers are far less random than you would think.

Computers have the same problem. Although all systems are able to pick random numbers, the method they use is often tied to specific other numbers -- for example, the time -- that means you could get a very similar series of 'random' numbers in different situations.

This tendency manifests itself in many ways. For anyone who uses their iPod heavily, you've probably noticed that your supposedly random 'shuffling' iPod seems to be particularly fond of the Bee Gees, Melissa Etheridge or Pavarotti. Look at a random playlist that iTunes generates for you, and you're likely to notice several songs from one or two artists, while other artists go completely unrepresented.

Source: Cnet

Probability

- There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.
  - $P(A)$ = Probability of event A
  - $0 \leq P(A) \leq 1$

Probability

- There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.
  - $P(A)$ = Probability of event A
  - $0 \leq P(A) \leq 1$
- ***Frequentist interpretation:***
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability

- There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.
  - $P(A)$ = Probability of event A
  - $0 \leq P(A) \leq 1$
- ***Frequentist interpretation:***
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
- ***Bayesian interpretation:***
  - A Bayesian interprets probability as a subjective degree of belief: For the same event, two separate people could have different viewpoints and so assign different probabilities.
  - Largely popularized by revolutionary advance in computational technology and methods during the last twenty years.

Law of large numbers

Law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome, $\hat{p}_n$, converges to the probability of that outcome, p.

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$H\ H\ H\ H\ H\ H\ H\ H\ H\ H\ ?$$

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.

$$P(H \text{ on } 11^{th} \text{ toss}) = P(T \text{ on } 11^{th} \text{ toss}) = 0.5$$

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.

$$P(H \text{ on } 11^{th} \text{ toss}) = P(T \text{ on } 11^{th} \text{ toss}) = 0.5$$

- The coin is not "due" for a tail.

Law of large numbers (cont.)

When tossing a *fair* coin, if heads comes up on each of the first 10 tosses, what do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- The probability is still 0.5, or there is still a 50% chance that another head will come up on the next toss.

$$P(H \text{ on } 11^{th} \text{ toss}) = P(T \text{ on } 11^{th} \text{ toss}) = 0.5$$

- The coin is not "due" for a tail.
- The common misunderstanding of the LLN is that random processes are supposed to compensate for whatever happened in the past; this is just not true and is also called ***gambler's fallacy*** (or ***law of averages***).