# Chapter 5: The normal approximation for data

### Context

- We still look at summarizing data
- Many histograms follow the normal curve, if they are drawn in standard units
- So we'll discuss what standard units are
- And you'll learn to work with the normal curve
- Then we'll combine these two. We will use the normal curve to estimate the percentage of entries of a real data set in a given interval
- Other new subjects: percentiles, percentile-ranks, box-plot, interquartile range

2 / 21

## Normal curve

3 / 21

### Normal curve

- Other names: Gaussian curve, bell curve
- Discovered in 1720 by Abraham de Moivre
- See picture on overhead
- Important properties:
  - the curve is symmetric about 0
  - the total area under the curve is $100\%$
  - the curve is always above the horizontal axis

4 / 21

### Areas under the normal curve

- Areas under the normal curve to memorize:
  - Area between -1 and 1: $68\%$
  - Area between -2 and 2: $95\%$
  - Area between -3 and 3: $99\%$
- We can find other areas under the curve using:
  - the table in the back of the book
  - total area under the curve is $100\%$
  - curve is symmetric about zero
- How to do this?
  - sketch normal curve and shade area of interest
  - build the area from areas that you know from the table
- See examples on overhead (Section 5.2, Exercise set B)



## Normal approximation



### Standard units

- Why use standard units?
  - Many histograms follow the normal curve if they are drawn in standard units
- How to convert a value to standard units?
  - Compute how many SDs the value is above or below average. Values above average have a plus sign, values below average have a minus sign.



### Examples

- Example: women in HANES study. Average height is 63.5 inches, SD is 2.5 inches. Convert the following heights to standard units.
  - 68.5 inches. Answer: 68.5 inches is 5 inches above the average. That is two SDs above the average.Her height in standard units is 2.
  - 61 inches. Answer: 61 inches is 2.5 inches below the average. That is one SD below the average. Her height in standard units is -1.
  - 63.5 inches. Answer: 63.5 inches is right at the average. So this value is zero SDs above the average. Her height in standard units is 0.
  - See Figure 2 on page 81

## 68% - 95% rules

- 68% rule: for many lists 68% of the entries are between average - SD and average + SD. Where does this rule come from?
  - convert the interval to standard units: -1 to 1
  - the area under the normal curve between -1 and 1 is 68%
  - if the histogram follows the normal curve, the area under the histogram is about 68%
- 95% rule: for many lists 95% of the entries are between average - 2 SDs and average + 2 SDs. Where does this rule come from?
  - convert the interval to standard units: -2 to 2
  - the area under the normal curve between -2 and 2 is 95%
  - if the histogram follows the normal curve, the area under the histogram is about 95%



## Normal approximation for data

- We have histogram that follows that normal curve. We know the average and SD. We want to find the percentage of entries in a certain interval.
- Solution:
  - Step 1: Draw the number line and shade the interval of interest. Mark the average on the line.
  - Step 2: Convert to standard units. Draw a new line in standard units, and shade the interval of interest.
  - Step 3: Sketch the normal curve, and find the area above the shaded interval of Step 2.
- See examples on overhead (Section 5.2, Exercise set C)

## Percentiles and percentile ranks

### Percentiles

- We can also summarize data by looking at percentiles. What are percentiles?
- 1st percentile = number such that $1\%$ of the entries are smaller than the number, and $99\%$ are larger
- 10th percentile = number such that $10\%$ of the entries are smaller than the number, and $90\%$ are larger
- 25th percentile = number such that $25\%$ of the entries are smaller than the number, and $75\%$ are larger
- 50th percentile = number such that $50\%$ of the entries are smaller than the number, and $50\%$ are larger
- 75th percentile = number such that $75\%$ of the entries are smaller than the number, and $25\%$ are larger
- etc.



### Quartiles

- 1st quartile = number such that one quarter of the entries are smaller than that number, and three quarters are larger = 25th percentile
- 2nd quartile = number such that two quarters of the data are smaller than that number, and two quarters are larger = 50th percentile = median
- 3rd quartile = number such that three quarters of the entries are smaller than that number, and one quarter is larger = 75th percentile
- interquartile range = 3rd quartile - 1st quartile
- See summary of homework grades, and example with Table 1 on page 89



### Percentile rank

- The percentile rank of a value is the percentage of entries smaller than that value.
- Examples:
  - the percentile rank of the highest homework score is $100\%$
  - the percentile rank of the median homework score is $50\%$

### Percentile vs. percentile rank

- Percentile and percentile rank are each other's opposites
- Example income data:
  - The 10th percentile is $\$10,200$. It is the number such that $10\%$ of the entries are smaller than that number.
  - The percentile rank of $\$10,200$ is $10\%$. It is the percentage of entries smaller than $\$10,200$.
  - Yet another way of saying this: An annual income of $\$10,200$ puts you at the 10th percentile of the income distribution
- Remember:
  - A percentile is a number
  - A percentile rank is a percentage



### Percentiles and the normal curve

- If a histogram follows the normal curve, then the normal curve can be used to estimate percentiles
- Method:
  - sketch a normal curve; find the right value of $z$, using the normal table
  - $z$ is given in standard units; convert it back to the units in the problem
- See examples (Section 5.5, Exercise set E)



## Change of scale



### Motivation

- Changes of scale often occur:
  - length: cm - inches - feet
  - temperature: Fahrenheit - Celsius
- Such changes of scale consist of:
  - multiplying all entries by a constant
  - adding a constant to all entries
  - or both of the above
- What do such changes do to the average, SD and standard units?
- See examples on overhead

**Effects of change of scale**

- If we add a constant to all entries in a list, then
  - the average increases by this constant
  - the SD does not change
  - the standard units do not change
- If we multiply all entries in a list by a positive number, then
  - the average is multiplied by this number
  - the SD is multiplied by this number
  - the standard units do not change

19 / 21

## Data that don't follow the normal curve

20 / 21

**How to summarize data that don't follow the normal curve?**

- Don't use the normal curve!
- Good ways to summarize such data are:
  - the histogram
  - a table with percentiles (like for the income data)
  - the 1st quartile, median and 3rd quartile (like for the homework grades)
  - box-plot:
    - box given by 1st and 3rd quartile: contains the middle $50\%$ of the data
    - median is given as a line in the box
    - it also gives some information on entries that fall outside the box
    - see example for homework grades

21 / 21