Math 3338: Probability (Fall 2006)

Jiwen He

Section Number: 10853

`http://math.uh.edu/~jiwenhe/math3338fall06.html`

Course Information

- **Classes:** 11:00-12:00 MWF in 140SR
- **Instructor:** Dr. Jiwen He
  - Office: 684 PGH
  - Phone: (713) 743-3481
  - E-mail: jiwenhe@math.uh.edu
  - Office hours: 12:00-13:00 MW or by appointment

  Textsbook: *Modern Mathematical Statistics with Applications*, by J.L. Devore and K.N. Berk, Duxbury, 2007
- **Prerequisites** Math 1432 (Calculus II)
- **Course Homepage:**

  `http://math.uh.edu/~jiwenhe/math3338fall06.html`
- **Extra help:** Mathlab, 222 Garrison Gym: 9 AM - 7 PM Mo-Th, 9 AM - 2 PM Fr

Course Policies and Procedures

Grading

- There will be homeworks, four mid-exams and a final exam. The final is comprehensive and compulsory. The course grade is approximately based on a total score of 800 points (150 points for the homeworks, 100 points for each mid-exam, 250 points for the final).
- There will be no make-ups for the missed exams, unless there is a very serious reason. In such a situation, I expect to be notified as soon as it is reasonably possible.
- Students with a valid excuse for missing up to one mid-exam must provide written documentation to that effect, e.g., a medical certificate.
- If you miss more than one mid-exam or the final, you will either be dropped by the instructor or get an "F" as the final grade of this course.
- Late homeworks (up to one class period) will receive a 20% penalty.
- The grade A will not be given to a person with the lower homework assignment score.
- You can drop with a W until the last day to drop the course. Incompletes are given only in unusual circumstances, and never just to prevent a bad grade.

Mid-Exams

- The mid-exam will be given on Sep 13, Sep 25, Oct 20, Nov 13 in class (40 minutes).
- It will cover all material up to what was discussed two classes before the mid-exam
- Please bring your student I.D.
- It will be a closed book exam. The necessary tables will be provided. However, you can bring one sheet of paper (regular size, two-sided) with notes. You can write whatever you want in these notes, but it should be your own work.
- Calculators can be used.
- Review will be given during the class before the exam. Bring any questions you might have.

Final Exam

- It will be on Wed, Dec 13, 11 am-2 pm in 140SR,and will be comprehensive.
- Please bring your student I.D.
- It will be a closed book exam. The necessary tables will be provided. However, you can bring two sheets of paper (regular size, two-sided) with notes. You can write whatever you want in these notes, but it should be your own work.
- Calculators can be used.
- Review of the materials discussed this semester will be given on Dec. 1.

Homework

- Homework problems will be assigned each week.
- Homeworks are due Wednesday, and will be returned the following Monday.
- Homework submitted up to one class period (not one week) beyond their due date will receive a 20% penalty.
- Homework submitted later than one class period beyond its due date will not be accepted without a written excuse.
- Homework scores can not be changed one week after they have been returned.
- In order to learn the material and do well on exams you should solve the assigned problems. Try to work the problems by yourself, or together with your colleagues. You are encouraged to form study groups. If you have difficulties, ask questions in class (this helps everyone).
- You are encouraged to discuss homework with your classmates. However, you are expected to individually write up your solutions.
- Please also take advantage of the office hours.

Chapter One
Overview and Descriptive Statistics

1.1 Populations and samples

Definitions

- **Population:** a well-defined collection of objects.
- **Census:** desired information is available for all objects in population.
- **Sample:** a subset of the population.
- **Variable:** is any characteristic whose value may change from one object to another in the population, denoted by lowercase letters x, y, z.
  - **Univariate data:** observations on a single variable.
  - **Bivariate data:** observations on each of two variables.
  - **Multivariate data:** observations on more than two variables

Branches of Statistics

- **Descriptive statistics:** methods to summarize and describe important features of data.
  - *Graphical methods:* the constructions of histogram, boxplots, and scatter plots.
  - *Numerical methods:* the calculations of mean, standard deviations, and correlations coefficients.
  - *Computer software packages:* MINITAB, SAS, and S-Plus.
- **Inferential statistics:** methods for generalizing from a sample to a population.
  - *Inferential procedures:* point estimation, hypothesis testing, and estimation by confidence intervals.
  - *Chapters 7-14:* the subject of the 2nd semeter statistics

Example 1.1. Challenger accident

```
Stem-and-leaf of temp N = 36
Leaf Unit = 1.0
 1    3   1
 1    3
 2    4   0
 4    4   59
 6    5   23
 9    5   788
13    6   0113
(7)   6   6777789
16    7   000023
10    7   556689
 4    8   0134
```

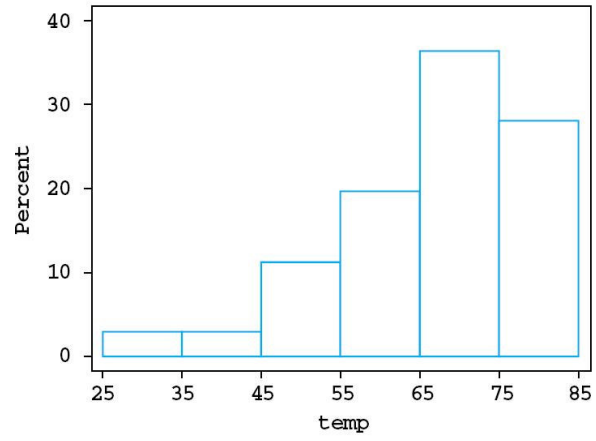

Figure 1.1 A MINITAB stem-and-leaf display and histogram of the O-ring temperature data

- **Data:** observations on x= O-Ring temperature for each firing
- **Graphs (Descriptive statistics):** A MINITAB stem-and-leaf display and histogram of the data: how the data is distributed
- **Features:** the lowest temperature is 31 degrees, much lower than the next-lowest temperatur, and this is the observation for the Challenge disaster.
- **Analysis:** warm temperatures are needed for successful operation of the O-rings.

Example 1.2. Elementary school students IQ

- **Sample**: 33 first-grade students in an elementary school
- **Data:** the IQ scores of the 33 students

82 96 99 102 103 103 106 107 108 108 108 108 109 110 110 111 113
113 113 113 115 115 118 118 119 121 122 122 127 132 136 140 146

- **Method (Inferential statistics):** estimate of the population mean IQ for the first grader: between 109.2 and 118.2 - a confidence interval.
- **analysis:** this is an above average class, the nationwide IQ average is around 100

Probability vs. Inferential statistics

- **Probability:** a bridge between the descriptive and inferential techniques

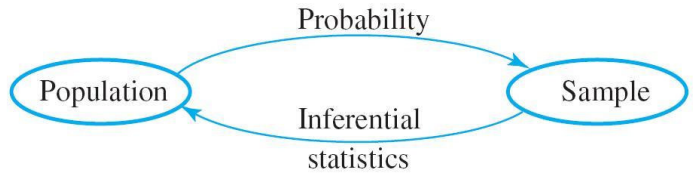

Figure 1.2 The relationship between probability and inferential statistics

- **Relations:**
  - Probability and inferential statistics both deal with questions involving populations and samples, but do so in an "inverse manner" to one another.
  - Before we can understand what a particular sample can tell us about the popultion, we should first understand the uncertainty associated with taking a sample from a given population. This is why we study probability before statistics.
- **Example 1.3 - Seat belts:** Contrasting focus of probability and inferential statistics
  - *Probability*: how many of the driver in a sample of size 100 can we expect to regularly use their lap belt?
  - *Inferential statistics:* a sample of 100 drivers revealed that 65 regularly use their lap belt. Does this provide substantial evidence for concluding that more than 50% of all drivers in this area regularly use their lap belt?

1.2 Pictorial and tabular methods in descriptive statistics

Stem-and-Leaf Displays

- **Steps for constructing a Stem-and-Leaf display:**
  1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
  2. List possible stem values in a vertical column.
  3. Record the leaf for every observation beside the corresponding stem value.
  4. Indicate the units for stems and leaves someplace in the display.
- **Information:** A stem-and-leaf display conveys information about the following aspects of the data:
  - Identification of a typical or representative value
  - Extent of spread about the typical value
  - Presence of any gaps in the data
  - Extent of symmetry in the distribution of values
  - Number and location of peaks
  - Presence of any outlying values>
  - Idendification of a typical or representative value

Example 1.4: Drinking on Campuses

0	4
1	1345678889
2	1223456666777889999
3	011223334455566667777888899999
4	1112222233444455666666777888888999
5	00111222233455666667777888899
6	01111244455666778

Stem: tens digit
Leaf: ones digit

Figure 1.3 Stem-and-leaf display for percentage binge drinkers at each of 140 colleges

- A typical or representative value is in the stem 4 row, perhaps in the mid-40% range.
- The observations are not highly concentrated about this typical value.
- The display rises to a single peak as we move downward, and then declines; there are no gaps in the display.
- The shape of the display is not perfectly symmetric, but instead appears to stretch out a bit more in the direction of low leaves than in the direction of high leaves.
- Lastly, there are no observations that are unusually far from the bulk of the data (no outliers), as would be the case if one of the 26% values had instead been 86%.
- The most surprising feature of this data is that, at most colleges in the sample, at least one-quarter of the students are binge drinkers.

Example 1.5: Golf Courses

Stem: Thousands and hundreds digits
Leaf: Tens and ones digits

64	35 64 33 70
65	26 27 06 83
66	05 94 14
67	90 70 00 98 70 45 13
68	90 70 73 50
69	00 27 36 04
70	51 05 11 40 50 22
71	31 69 68 05 13 65
72	80 09

(a)

Stem-and-leaf of yardage N = 40
Leaf Unit = 10

4	64	3367
8	65	0228
11	66	019
18	67	0147799
(4)	68	5779
18	69	0023
14	70	012455
8	71	013666
2	72	08

(b)

Figure 1.4 Stem-and-leaf displays of golf course yardages: (a) two-digit leaves; (b) display from MINITAB with truncated one-digit leaves

- Among the sample of 40 courses, the shortest is 6433 yards long, and the longest is 7280 yards.
- The lengths appear to be distributed in a roughly uniform fashion over the range of values in the sample.
- A stem choice here of either a single digit (6 or 7) or three digits (643, ..., 728) would yield an uninformative display, the first because of too few stems and the latter because of too many.

Dotplot

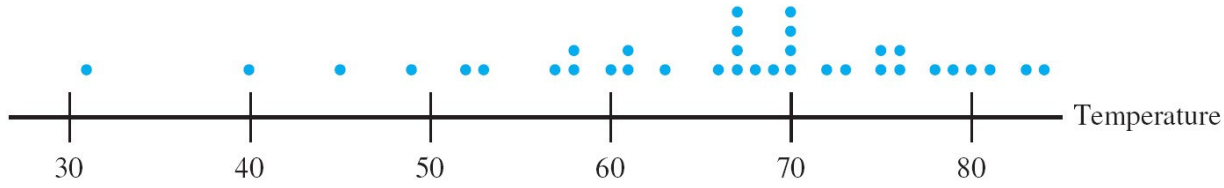

Figure 1.5 A dotplot of the O-ring temperature data (°F)

- A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values.
- Each observation is represented by a dot above the corresponding location on a horizontal measurement scale.
- When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.
- As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

HISTOGRAM

- **Definitions:**
  - The **frequency** of any particular value is the number of times that value occurs in the data set.
  - $$\text{The } \textbf{relative frequency} \text{ of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}.$$
  - A frequency distribution is a tabulation of the frequencies and/or relative frequencies.
- A Histogram for counting data:
  1. Determine the frequency and relative frequency of each x value.
  2. Mark possible x values on a horizontal scale.
  3. Above each value, draw a rectangle whose height is the relative frequency (or alternatively, the frequency) of that value.

Example 1.7: Baseball Games

Table 1.1 Frequency distribution for hits in nine-inning games

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

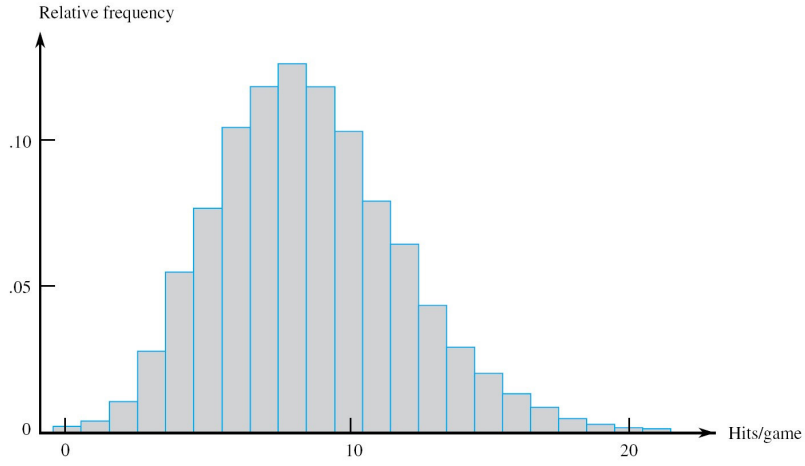

Figure 1.6 Histogram of number of hits per nine-inning game

proportion of games with at most two hits

$$= 0.0010 + 0.037 + 0.0108 = 0.0155$$

proportion of games with between 5 and 10 hits

$$= 0.0752 + \cdots + 0.1015 = 0.6361$$

Example 1.8: Energy Consumption

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

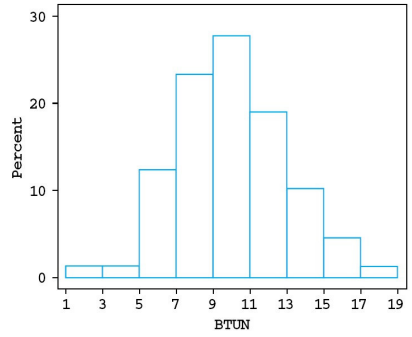

Figure 1.7 Histogram of the energy consumption data from Example 1.8

Class	1–3	3–5	5–7	7–9	9–11	11–13	13–15	15–17	17–19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011

Constructing a histogram for measurement data entails subdividing the measurement axis into a suitable number of class intervals or classes, such that each observation is contained in exactly one class.

- **A histogram for data with equal class widths:**
  1. Determine the frequency and relative frequency for each class.
  2. Mark the class boundaries on a horizontal measurement axis.
  3. Above each class interval draw a rectangle whose height is the corresponding relative frequency.

Example 1.9: Bonding Strength

Class	2–4	4–6	6–8	8–12	12–20	20–30
Frequency	9	15	5	9	8	2
Relative frequency	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

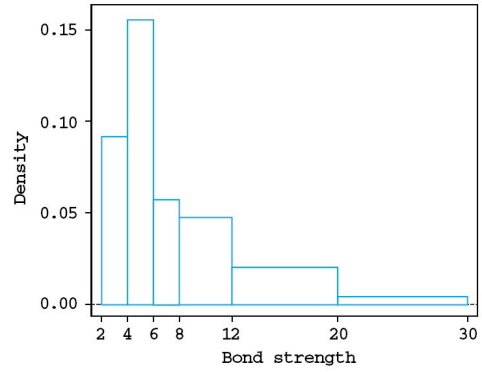

Figure 1.9 A MINITAB density histogram for the bond strength data of Example 1.9

- **A histogram for data with unequal class widths:** After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

- The resulting rectangle heights are usually called **densities**, and the vertical scale is the **density scale**. This prescription will also work when class widths are equal.

Histograms Shapes

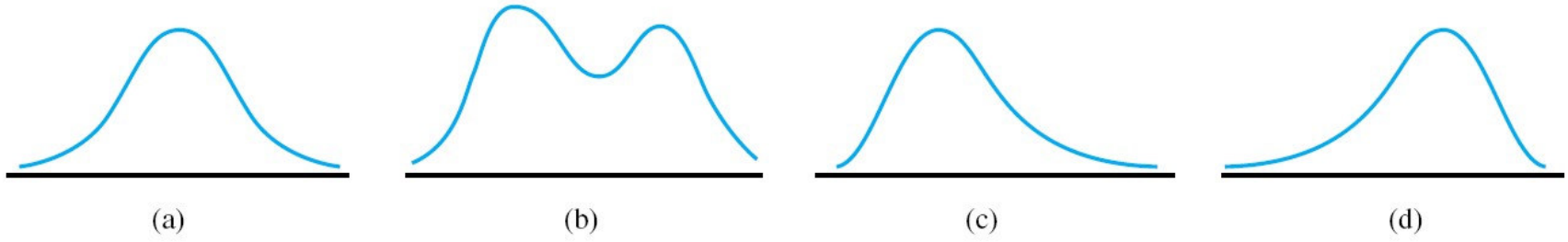

Figure 1.10 Smoothed histograms: (a) symmetric unimodal; (b) bimodal; (c) positively skewed; and (d) negatively skewed

- Histograms come in a variety of shapes:
  - A **unimodal** histogram is one that rises to a single peak and then declines.
  - A **bimodal** histogram has two different peaks.
  - A histogram with more than two peaks is said to be **multimodal**
- A histogram is **symmetric** if the left half is a mirror image of the right half.
- A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail and **negatively skewed** if the stretching is to the left.
- Figure 1.10 shows smoothed histograms, obtained by superimposing a smooth curve on the rectangles, that illustrate the various possibilities.