Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 2: April 18, 2013

Abstract

Review discussion of controlled experiments and observational studies. Present specific example of apparent gender bias in graduate school admissions. Introduce distributions and their presentations (density and histogram).

About Experiments and Observation

- Statistical analysis is *art*, not science.
- We often have a choice of *what* data we collect, and how to collect it.
- In a *controlled experiment* we have substantial control over the *relationship* among observations and their *quantity*, and we exploit that control to achieve a clear comparison among results in different circumstances.
- Otherwise, we are conducting an *observational study*.

Experiment or Observation?

- Other things being equal, *experimental evidence is preferred* because it is better controlled and provides clearer contrast between acceptance and rejection of a hypothesis.
- Nevertheless an observational study may be preferred for *ethical*, *financial*, or *feasibility* reasons.
- It is always useful to have more evidence because of uncertainty. Therefore, although we consider experimental evidence more reliable, we admit pure observation when available, and are dissatisfied with a theory that predicts experimental outcomes well but cannot explain other observations.
  - The Salk vaccine field trial provides an example.

The Salk vaccine field trial

- The ethical problem ("if it's safe and might work, shouldn't all children be vaccinated now?") was resolved by considering the great expense of a nationwide program.
  - Note that many parents refused, anyway.
- Two different designs were proposed.

  NFIP *Arbitrarily* split sample into *control group* (1st and 3rd graders) and *experimental group* (2nd graders). Parental permission was requested for experimental group; some refused.

  Experts group Requested permission for whole sample; some parents refused. No-permission students were excluded from experiment. The rest were randomly assigned to experimental or control group.

Analysis of Salk Field trials

- *Experts group* design: better *in theory*, clearer results *in practice*.
- No-permission students in experts group design were clearly different from either controls or experimental group, so excluding them from experiment was correct.
  - This is one way to use observational evidence, referred to above. Note that this difference helps to explain why the NFIP design has not very good, because it shows *how* the NFIP design lost some control (parents help decide who goes in which group).
- The fact that both studies were conducted is a *natural experiment*. It was not done intentionally, but from the point of view of analyzing the effects of different statistical methods it's hard to see how a controlled experiment could be done better.

Other studies and implications

- Several other studies on surgery and drug treatments were mentioned (and are described in detail in the Freedman *et al.* text).
- These show that medical studies are often biased in favor of new treatments, and that use of randomization (and double-blind: doctors are experimental subjects, too, not just the patients!) helps.
- For surgery, the mechanism of bias is clear: doctors only do surgery on relatively healthy patients. If non-candidates are used as controls, they have naturally lower survival rates than candidates, biasing the results in favor of surgery.

Various difficulties in design

- **Principle:** relationships among observations and variables must be controlled. Ideally, the only difference between experimental and control groups is the treatment. What can go wrong?
- *Imbalanced sample: counts* will also be unbalanced, biasing results toward large group.
- *Self-selection: rates* will be biased if the choice criteria (wealth of parents) are related to the variable of interest (infection by polio).
- *Confounding: rates* will be biased if some unknown factor associates the outcome with *treatment*.
- *Placebo effect:* the treatment causes an effect through other channels.

Addressing the difficulties

- *Imbalanced sample:* computing *rates* removes the bias.
- *Self-selection:* imposing treatment *after* self-selection.
- *Confounding:* assign treatments *randomly*.
- *Placebo effect:* use a *placebo* on the controls, and *double-blind* on those who evaluate the effect.

Observational studies

- In controlled experiments, the investigators decide who is in the control group, and who is in the treatment group.
- In observational studies, the investigators have no choice. Treatment may be determined by
  - Self-selection by the subjects
  - Random selection by a natural process
  - Systematic selection by some process (including a different experimental study!)
- Statistical techniques to control for selection differ substantially depending on the selection process.

All this "control"

- Scientific terminology is often confusing. So far the word "control" means three different things.
  - *Controlled* experiment means choosing which subjects get which treatments
  - A *control* in an experiment is a subject who gets no special treatment
  - *Controlling* for confounding factors in an observational study means subtracting their effects out by statistical methods

Gender bias?

	Men		Women	
Major	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
		44%		30%

Why the gender differential?

- The last line of the table is absolutely clear: on average, a male applicant is 50% more likely to be admitted than a female one.
- Can this be attributed to gender bias on the part of admission committees? **No:** In no major are men as much as 20% more likely to be admitted, and in half of the majors women are *more likely* to be admitted, sometimes dramatically so.
- There *is* a gender bias: the majors that men are likely to apply to have a higher admission rate than those that women are more likely to apply to. But this is choice by the "victims," not discrimination by the faculty.

Lack of control leads to bias

- Based on the admission rates by major and gender, there is no visible discrimination (except in major A, in favor of *women*!)
- The feature of the study that leads to confusion is *lack of control of who gets which treatment* (*Treatment* is "which major to apply to," *not* the *outcome* of "admitted or not").
- Major determines admission rate, but gender determines major. The *indirect* cause and effect relationship between gender and admission rate (via choice of major) is *confounded* with the *direct* causal relationship of major to admission rate.
- Frequency of *mediating variables* (like choice of major) is crucial, but the effect is concealed by taking averages.

Controlling for confounding effects

- "Controlling the effect" of major choice requires an *experiment* in which applicants are not allowed to choose their own major, but instead are assigned a major to apply to.
  - Ethically unreasonable, practically infeasible.
- In an observational study, we "control *for* an effect" by collecting additional data that allows us to measure the confounding effect and "subtract" it from the observed total effect.

Organization of data

- Note that simply collecting the number of applicants to each major is not good enough. We need to also collect data on gender of applicants, and the outcomes.
- We can (and usually do) associate all the values (treatment variables, outcome variables, and confounding variables) for each individual into an *observation*.
- Computers make it possible to work with very large numbers of observations directly.
- However, for the calculations we need to do it is sufficient to have a *multivariate distribution*.

A multivariate distribution

- In the "gender bias?" table, there are *three variables* per observation (*i.e.*, applicant): *gender*, *major*, *decision*, for a total of $2 \times 6 \times 2 = 24$ *cells*.
  - In many cases each observation (*i.e.*, the set of all information we have on each candidate) will include other information (*e.g.*, name, address, university, and undergraduate major), but we can ignore those in computing the distribution of interest.
- In the "gender bias?" table, counts are only explicitly given for gender and major, but the counts for decision (*pass* or *fail*) for each gender-major group can be computed from the admission rate given.

A multivariate distribution: table

Gender	Major	Decision	**Count**
female	A	pass	89
female	A	fail	19
female	B	pass	17
female	B	fail	8
female	C	pass	202
female	C	fail	391
female	D	pass	131
female	C	fail	244
female	E	pass	94
female	E	fail	299
female	F	pass	24
female	F	fail	317

Gender	Major	Decision	**Count**
male	A	pass	512
male	A	fail	313
male	B	pass	353
male	B	fail	207
male	C	pass	120
male	C	fail	205
male	D	pass	138
male	C	fail	279
male	E	pass	53
male	E	fail	138
male	F	pass	22
male	F	fail	351

Golden Week schedule

- There is **no class** on May 5 due to Golden Week holidays.
- Drop-in **office hours are moved** from May 5 to Friday, May 7, at **13:30** (note: *time change!*)

Collection of observations (*data set*)

- In an *empirical study* (a study that collects and analyzes data) we measure the value (or for qualitative data, observe which of the possible values occurred) for *each individual to be measured.*
- Each individual's measurements are grouped into a single record, call an *observation*, for that individual.
- Looking carefully at the details of each individual is tedious, usually uninformative, and even infeasible for large data sets.
  - Picking selected *typical* or *atypical* individuals, and recording and analyzing many variables ("facts about them") can be very valuable and informative (this is the *case study method*).
  - Excessive detail about *many* individuals can be unprofitable.
- With large numbers of observations, use *frequency distributions.*

Frequency distribution

- A *frequency distribution* sorts the observations into different *cells* ("kinds"), according to their values of the variables, and measures the size of each group.
  - Alternatively, each possible value may be given a cell of its own, or added to larger and larger groups in a *cumulative frequency distribution*.
- The mapping from the cell to the size is the frequency distribution.
- Frequencies may be expressed as *counts*, *percentages*, or *densities*.

Frequency distribution examples

- Take a group of 25 students, and observe their midterm grades in statistics, for simplicity given as A, B, C, or D.
- One way to construct a distribution is to put each grade in its own cell:

Grade	D	C	B	A
Frequency	1	5	11	8

- The cumulative distribution is constructed by counting all the students with a given grade *or lower*:

Highest grade in cell	D	C	B	A
Cumulative frequency	1	6	17	25

Relative frequency distribution

- The corresponding *relative* (percentage) *distributions* are:

Grade	D	C	B	A
Relative frequency	4%	20%	44%	32%
Cumulative relative frequency	4%	24%	68%	100%

- Note that, whether absolute or relative, the cumulative frequency distribution is increasing, it is the sum of the "lower" cells in the frequency distribution, and it is bounded below by 0.
- In the relative case, it is bounded above by 1 (*i.e.*, 100%), and for the cumulative distribution, the cumulative frequency for the highest cell is always 100%.

Usage

- In empirical work, the most commonly used concepts are absolute and relative frequency distributions.
  - Absolute frequency distributions are useful to compare to absolute units in applications. For example, with a national income distribution, multiplying absolute frequency of in each cell by the corresponding income, and adding the product for all cells, gives gross domestic product.
  - Relative frequencies are easy to interpret as probabilities.
- Cumulative distributions are often expressed as *percentiles*. They also are technically useful in probability theory.
- *Densities* are used for histograms of continuous cardinal variables.

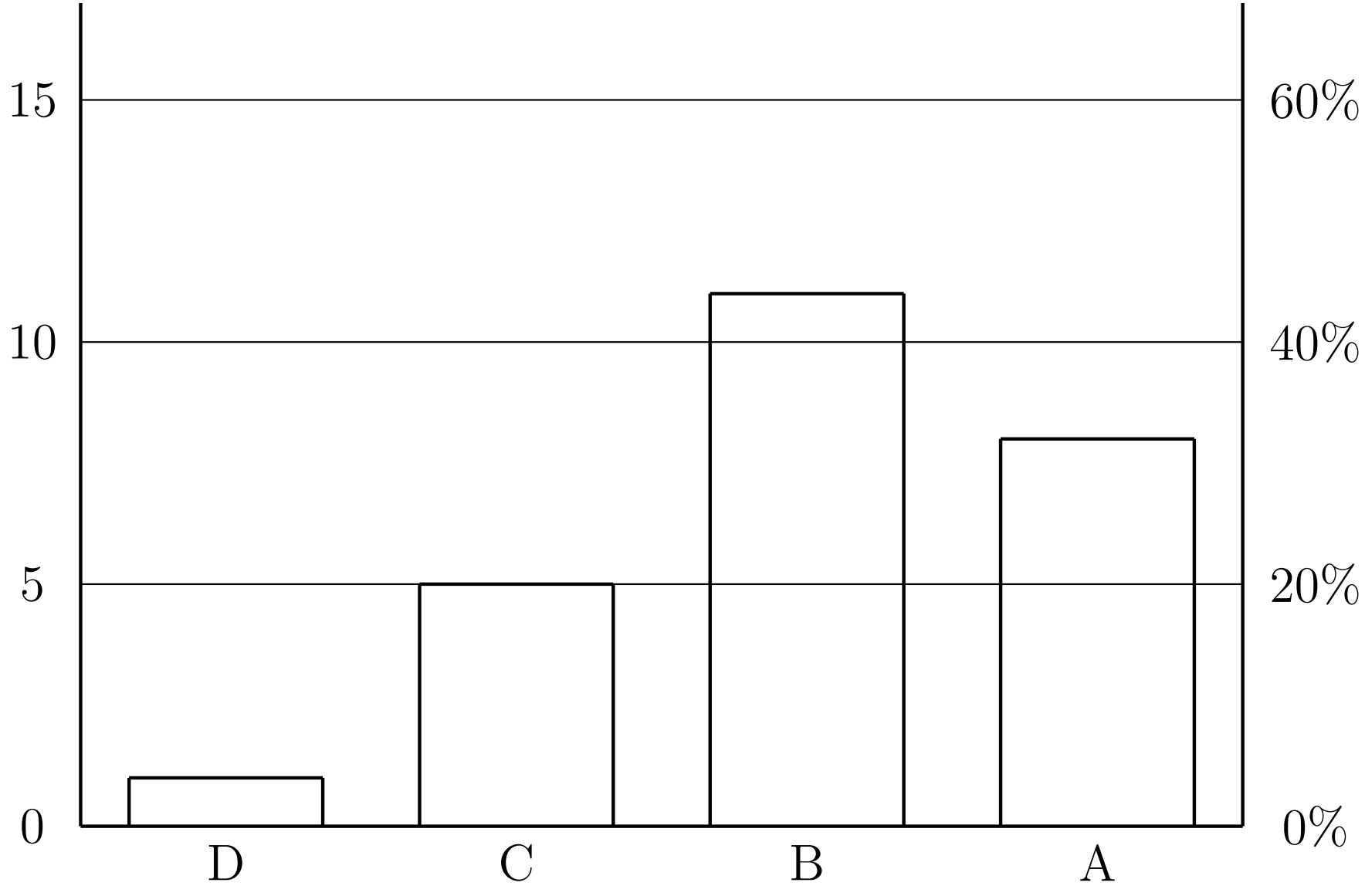
Be careful about terms!

- In empirical work both "frequency distribution" and "relative frequency distribution" are commonly abbreviated to just "distribution"
- In probability theory, "distribution" normally refers to the *cumulative distribution*, and the frequency distribution is called *(probability) mass function* or the *(probability) density function* depending on whether the distribution is discrete or continuous.
  - This usage is preferred because *distribution* is a general concept, but it is possible that neither a mass function nor a density function exists for a distribution, while the cumulative distribution always exists. (Such distributions are called *mixed distributions*.)

Histogram

- A *histogram* is a graphical representation of a frequency distribution.
  - Histograms are often assumed to be based on empirical data; the graph of a theoretical frequency distribution is typically just called a *graph*.
- The histogram has the *cell labels* on one axis and the *frequency* on the other.

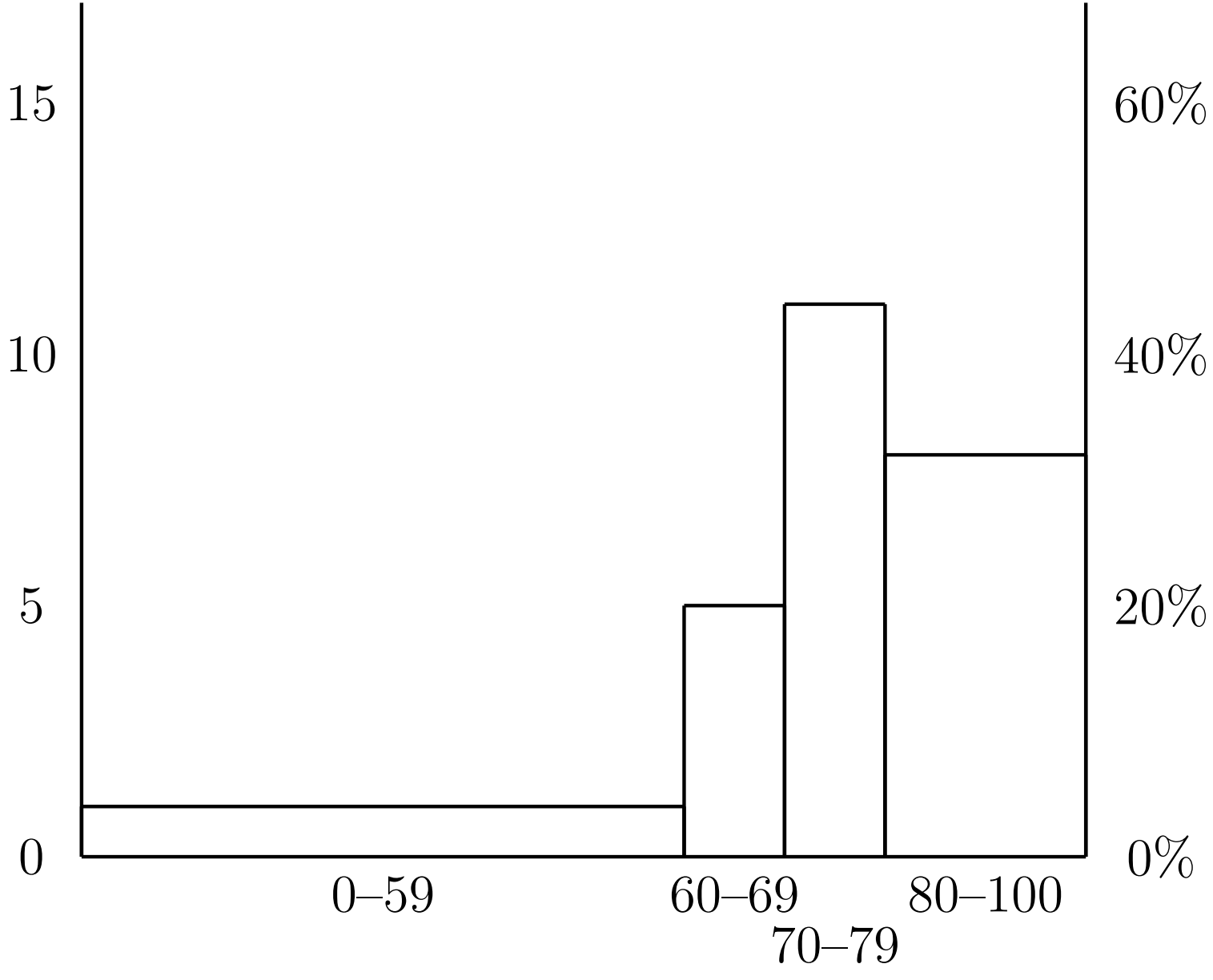
Histogram for the grade distribution



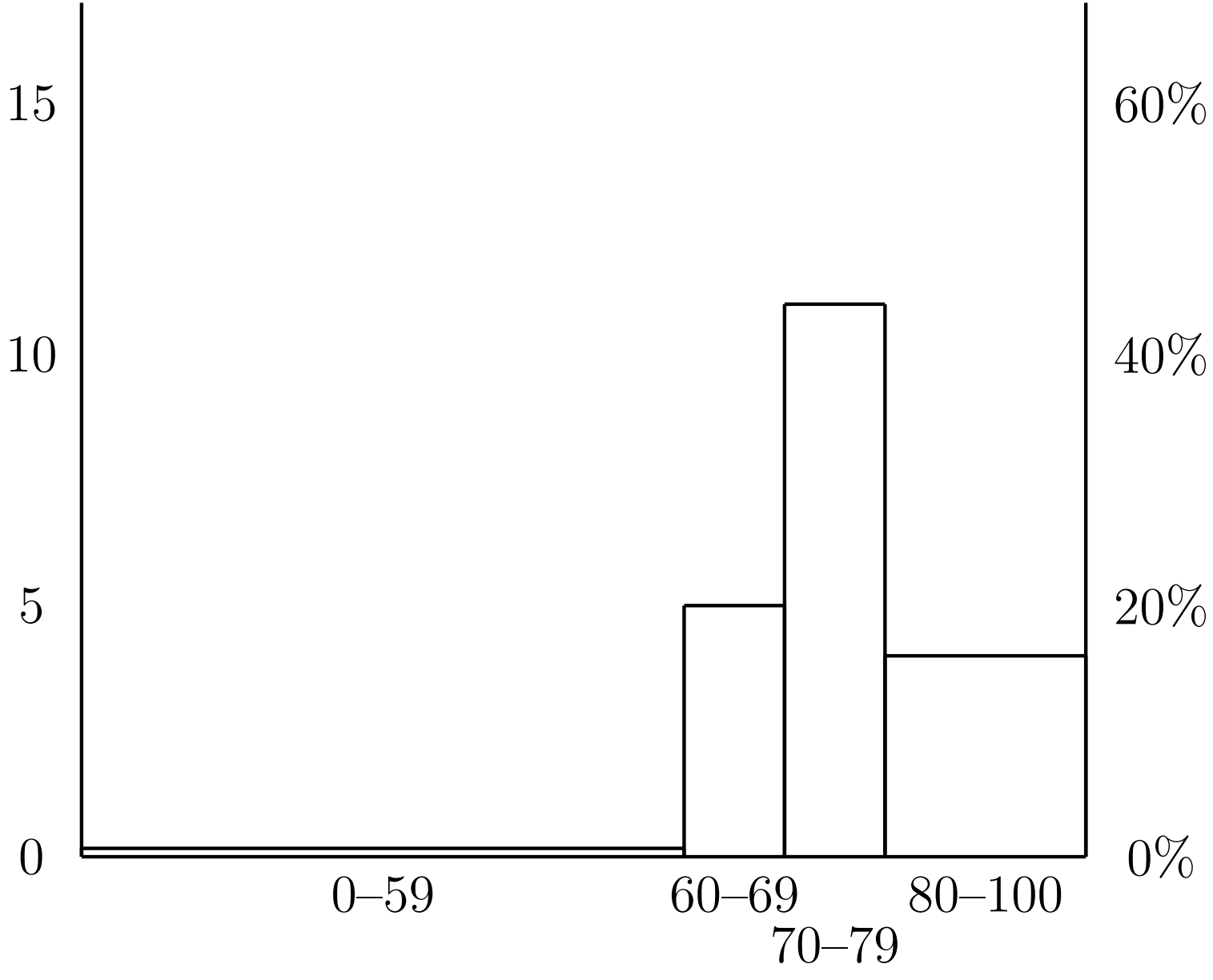
Densities

- Let's look at *raw scores* on the midterm, instead of letter grades.
- Assuming the usual translation $D < 60 < C < 70 < B < 80 < A$, we can define the cells 0–59, 60–69, 70–79, and 80–100.
- As fractions, the percentages are a *cardinal* scale. It makes sense to say "the range 80–100 is twice as wide as the range 70–79."
- If we rescale the horizontal axis to account for cardinality, the importance of the highest and lowest scores is overemphasized.
- So also rescale the vertical values. We don't know how many people fell into the ranges 80–89 and 90–100, but we *can* say that the average of the two cells of width 10 was 4. This is the *density* of the "combined cell" 80–100. The density of the combined cell 0–59 is 1/6.

Histogram without density adjustment



Histogram with density adjustment



Advantages of the density histogram

- The impression given by the two diagrams is quite different.
- The extremes (A and D) are *overemphasized* in the first histogram; the density histogram is more accurate.
- Two density histograms with *different cell definitions* can be compared; this is not possible without that adjustment.
- Relative density histograms can be directly compared with graphs of theoretical probability distributions.

Interpreting a density histogram

- Because of the density adjustment, frequency of a cell is not represented by height of the cell.
- Frequency of a cell is represented by its *area*.
- If cells have equal width, we can measure width in "cell-widths." Then the frequency is equal to the height.
  - To make the units work out, height must be measured in "frequency per cell-width."

Variable type and density

- The density adjustment can only be made for *cardinal* variables.
  - Otherwise "width of cell" is undefined!
- The density adjustment **must** be made for continuous variables.
  - It makes sense to speak of the frequency of a single value (say 1749 mm for the height of a man) in raw continuous data.
  - But you cannot draw a histogram with only the (finite number) of values observed ...a histogram **must** display frequency for all possible values. *Almost all* must be zero!
- So we smooth out a histogram for a continuous variable by using cells of non-zero width, giving a density.

Summarizing data: location

- Since there are many different values in a distribution, "location" must take account of all of them.
- For this to make sense, the variable values must be *ordered*, the variable must be *ordinal*.
  - There is an exception for the *mode*.

Location by the CDF

- The most certain way to compare the "locations" of two distributions is to use the cumulative relative frequency distribution.
  - All values of B may be greater than all values of A, obviously B as a whole is "higher" than A.
  - If A and B have the same number of observations, then we can pair them one-to-one, in order from lowest to highest. Every member of the A group is paired with a member of the B group which is at least as large.
- If *every* cumulative frequency is *smaller* in distribution B than in distribution A, then distribution B has a *higher* location.
  - B has *few low values*, so has *more high values*, and is "higher."

One-to-one comparison

- Consider two datasets of grades for two instances of the same class, with the same class size. For example:

  AM Class B C A A A B A C

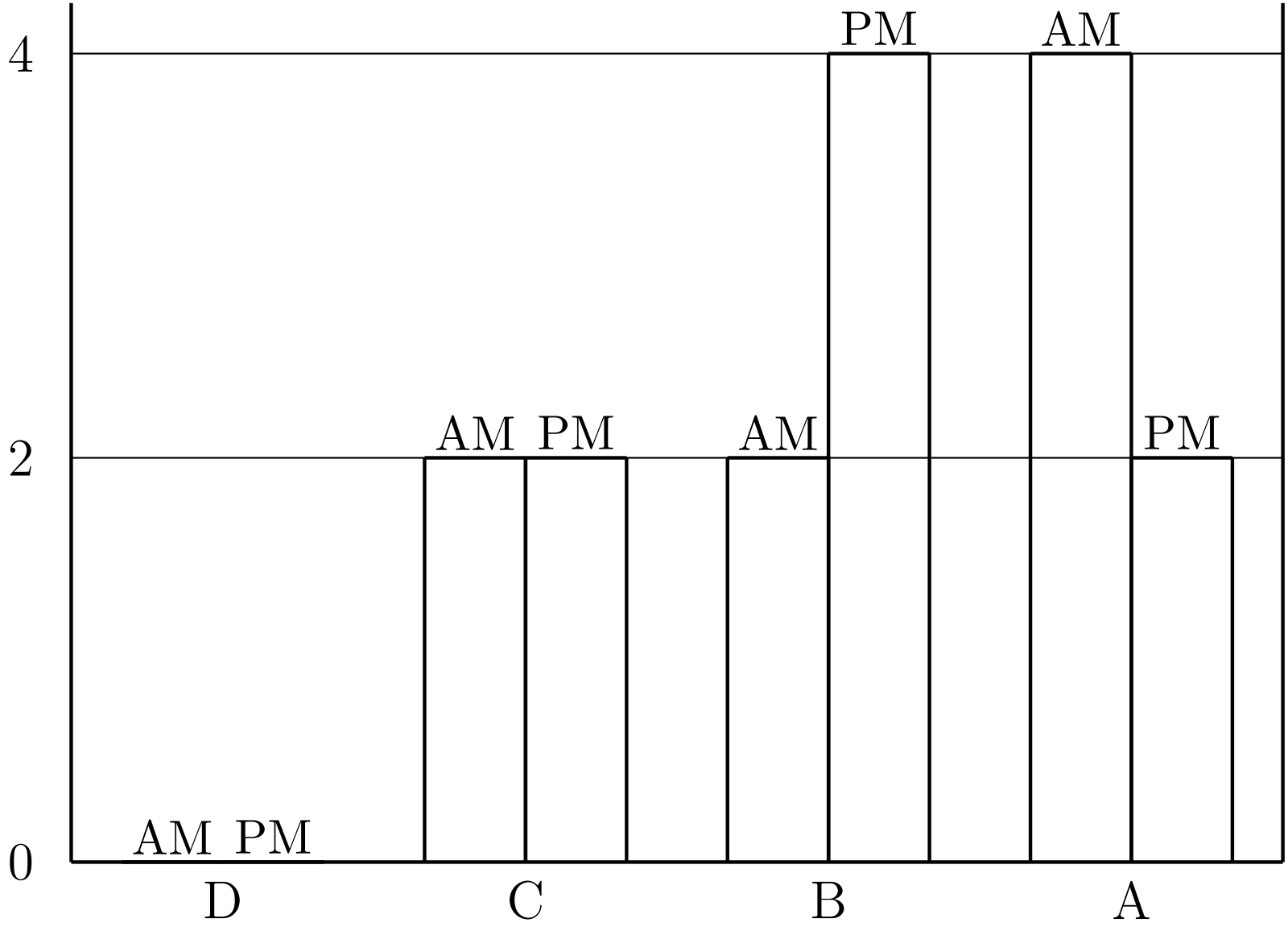
  PM Class A B C B B C A B

- Each class has As, Bs, and Cs. How can we say one class is better than another? By comparing the grade of the students at the same ranks. Easy because both classes are the same size.

AM Class	C	C	B	B	A	A	A	A
PM Class	C	C	B	B	B	B	A	A
"Winner"					AM	AM		

Table 1: Comparing data sets by rank order

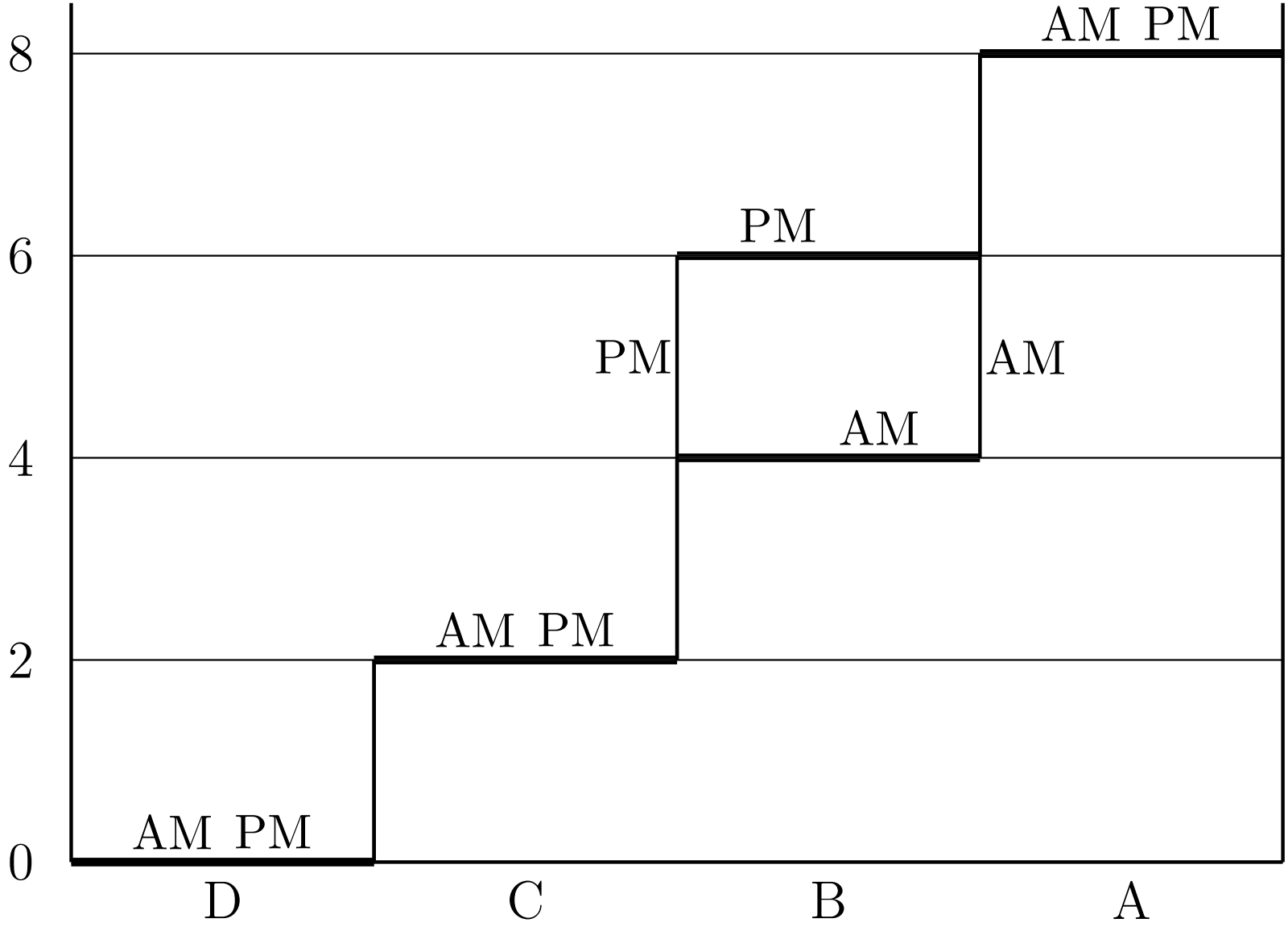
Comparing distributions



Comparing distributions

- The preceding graph displays two distributions on the same axes.
  - We use labels "AM" and "PM", but colors or shading are good.
  - They refer to *different data sets*.
- The comparison here is probably obvious to your eye, but what if the "spikes" were many, and of mixed heights?
- There is a way to make the cumulative comparison accurately using CDFs. If a CDF is to the right of another at all heights, it has higher location.
- The *edge* where the PM CDF rises above the AM CDF is between C and B, and the AM CDF "catches up" between B and A. Thus the #5 and #6 students have better grades in the AM class.

Comparing CDFs graphically



Homework 2: due April 25, 11:45am

Submit homework to `data-hw@turnbull.sk.tsukuba.ac.jp` by **email**. Note the due date is April 25, **11:45am**. Submission time is time of receipt by the server.

For this homework, please submit as *plain text* (no wordprocessor or PDF attachments). In other words, "just type" your answer in the email. In the *first line*, include your name, your student ID number, and the words "Homework 2".

You may answer in English or Japanese. In answering, you *must* explain why the answer you give is correct *in your own words* to receive credit.

If you wish to ask me questions and get an answer, *write a separate email* to `data-help@turnbull.sk.tsukuba.ac.jp`.

Problems

1. Is the questionnaire conducted in this class an *experiment* or an *observational study*?

Homework 3: due May 2, 2013 at 11:45am

Due 2013-05-02, 11:45 am.

Submit by email to `data-hw@turnbull.sk.tsukuba.ac.jp`. Your header should look like this:

```
From: a-student@sk.tsukuba.ac.jp
To: data-hw@turnbull.sk.tsukuba.ac.jp
Subject: Basic Data Analysis HW#3
```

The subject should be all half-width Roman letters (ASCII).

1. Download `homework-sheet-2.xls` from the home page. You also may want to get `mean-variance.xls` (used in class).
2. Copy the random numbers (values only!) to the *original data set* and construct the *sorted data set* as done in class.

3. Construct the absolute and relative frequency distributions, and the cumulative (relative) frequency distribution, of *letter grades*. (Space in the worksheet is allocated for the *dummy variable* technique shown in class, but if you prefer you can omit the dummy variables and use a count function or distribution function provided by the spreadsheet software.)

4. Translate the letter grades to the usual 4-point scale.

5. Enter the *mode* and the *median* in the appropriate places.

6. Enter formulæ to compute the various components of the *mean*, *variance*, and *standard deviation*. Also compute the variance using the "moment difference" method, and compare to the "central moment" method. They should be exactly the same.

7. In some empty space, draw a histogram of your distribution by using colored cells as shown in class.

8. Answer the questions following the distribution exercise in the spaces provided.

Important notes: For your convenience, areas provided for your answers are highlighted in blue.

In answering the questions, *you must explain your answers in some detail.* For numerical problems, you may provide the basic formulæ and detailed algebraic calculations (in general you should not show numerical calculations unless explicitly requested). For questions testing definitions, you should provide the definition and explain how the example corresponds to the conditions of the definition. For questions testing interpretation, you should describe the logic that leads to your conclusion.

There are a lot of questions on *skewness* and *kurtosis* in the problem set. You should not take this as an indication of the importance of these statistics themselves. Instead, the exercises are intended to

give you some "feeling" for the relationship between summary statistics and the "shape" of the distribution that they describe.

There are only two pure "interpretation" questions, no. 8 and no. 9. However, these are very important, and similar questions will be featured in both the midterm and the final examinations.