Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 3: April 25, 2013

Abstract

Review summary statistics and measures of location. Discuss the placement exam as an exercise in statistics. Describe measures of dispersion. Briefly mention higher moments (extending the idea of variance).

Midterm examination

- May 16, 12:15–13:30.
- Covers lecture material up to normal distribution and basic probability calculations.
  - Some past examinations are linked from the home page.
  - Study guide will be posted later.
- 4th period (13:45–15:00) **lecture will be conducted**.

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.*

Individuals

- An *individual* is the *unit of observation*, the “whole thing” that we measure, including:
  - quantitatively (height of a person), or
  - qualitatively (gender of a person);
  - objectively (the speed of a typhoon’s wind), or
  - subjectively (whether it’s a good day for a picnic).
- Each fact or measurement is the *value* of a *variable*.
- An individual can be any “thing” about which we can collect data: a person, a cow, a car, a star, a company, a nation....
- A *sample* is any set of individuals. A *population* is the set of *all* individuals.

Observations

- The collection of facts, judgments, and measurements of an individual is an *observation*.
- In statistics, we need to collect values of each variable for many individuals.
  - The ideal data set is "rectangular": we have a value for each variable for each individual.
- We also use the word *sample* for a data set on a sample of individuals. A data set for a *population* is called a *census*.

Domain and cells of a distribution

- A conceivable configuration of data for some individual is a *cell*. The configuration (values) of data differs from cell to cell, but all individuals in a cell have the same values for every variable.
- The *domain* of a distribution is the set of all possible cells. (The domain of the distribution of applicants to graduate school had 24 cells. The domain of a letter grade distribution has four cells.)
- The *support* of a distribution is the set of cells which actually contain some data. (The support of the applicant distribution was its domain, every cell contained individuals. The support of the AM/PM grade distribution was a subset of the domain, since no students got a D.)

Frequency distribution

- A *frequency distribution* sorts the observations into different cells.
- The mapping from the cell to the quantity of observations is the frequency distribution.
- Frequencies may be expressed as
  - *counts*: *absolute* frequency distribution
  - *percentages*: *relative* frequency distribution
  - *densities*: absolute or relative
- It is often useful to consider a *cumulative distribution.*

Cumulative distribution and density

- A cumulative distribution describes the total quantity less than or equal each possible value. Cumulative distributions can be constructed from absolute and relative frequency distributions, and from densities.
- In frequency distributions, the count or percentage of items in in the cell is measured. In densities, the count or percentage divided by the width of the cell is measured.
- For a cumulative distribution to make sense, the variable must be *ordinal*. For a density to make sense, the variable must be *cardinal*.

Measures of central tendency

- We saw that CDFs can be used to accurately say whether one distribution is “higher” than another. Three weaknesses:
  - Not summarized “enough”—too much data still.
  - Sometimes we need “absolute” location.
  - Sometimes the CDF comparison is ambiguous.
- Reduce the distribution to a single value, or *statistic*.
  - A *statistic* can be computed from the “original data” or from the corresponding distribution: the values *must be identical*.
  - Statistical computations use the values, but are “weighted” by the frequency of each value.
- Typical statistics of location are the *mode*, *median*, and *mean*.

The mode

- Have you ever eaten apple pie with ice cream? Then you have eaten “pie à la mode,” which is simply French for “following fashion.”

- In statistics, the *mode* is the most frequent value (or values), *i.e.*, the “most fashionable” value. As an equation, sometimes written $m = \arg\max f(x)$, where f is the frequency distribution.
  - The mode takes the same value whether you use the absolute or relative frequency distribution. Most statistics are computed with a relative frequency distribution.

Merits and demerits of the mode

- The mode is easy to calculate, or recognize from a histogram: it's the cell with the largest frequency.
- With "concentrated" distributions, the mode corresponds closely to the intuitive idea of "typical" or "representative."
- There may be several modes, which need not even be neighbors.
  - A distribution with more than one mode is called *multimodal*, with the common case of 2 modes getting the name *bimodal*.
- The mode ignores any values that are not modal.
  - Changing the rule for assigning members to cells can affect the mode dramatically.

Instability of the mode

- Because the mode does not take account of any values that are not modal, changing the rule for assigning members to cells can affect the mode dramatically.

- Consider a test with three "easy" all-or-nothing question worth 20 points each, and two "hard" questions with partial credit awarded in 1-point units.

- Suppose we divide a class of 20 students into 5 20-point cells: 0–20, 21–40, 41–60, 61–80, 81–100. A "typical" distribution might be 1, 3, 5, 7, 4. The mode is 61–80.

- However, suppose that the 3 students in the 21–40 range all got 2 of 3 easy questions, *i.e.*, 40 points, and no other score is repeated. If we take the distribution with 1-point cells, then the mode is 40!

The median

- The *median* is computed using the CDF. It is the *value* whose cumulative relative frequency is 1/2.
  - If the data values (including repeats!) are sorted in order of value, it is the middle value.
- In discrete distributions, there is typically not a single value with a cumulative distribution of exactly 1/2.
  - The median is the cell whose "rising edge" intersects the 1/2 line (*i.e.*, the smallest cumulative frequency greater than 1/2).
  - If there *is* a single value, then actually the median is *between* that cell and the next *higher* one.
- It is easiest to do this computation with a table of the CDF.

Recall the morning/afternoon classes

- The grade data:

AM Class	B	C	A	A	A	B	A	C
PM Class	A	B	C	B	B	C	A	B

- Sorting ("ranking") the grades in each class:

AM Class	C	C	B	B	A	A	A	A
PM Class	C	C	B	B	B	B	A	A

- The absolute frequency distributions:

Grade	D	C	B	A
AM Class	0	2	2	4
PM Class	0	2	4	2

The median and the CDF

- Looking at our morning and afternoon classes' grade distributions again, we compute the absolute CDFs:

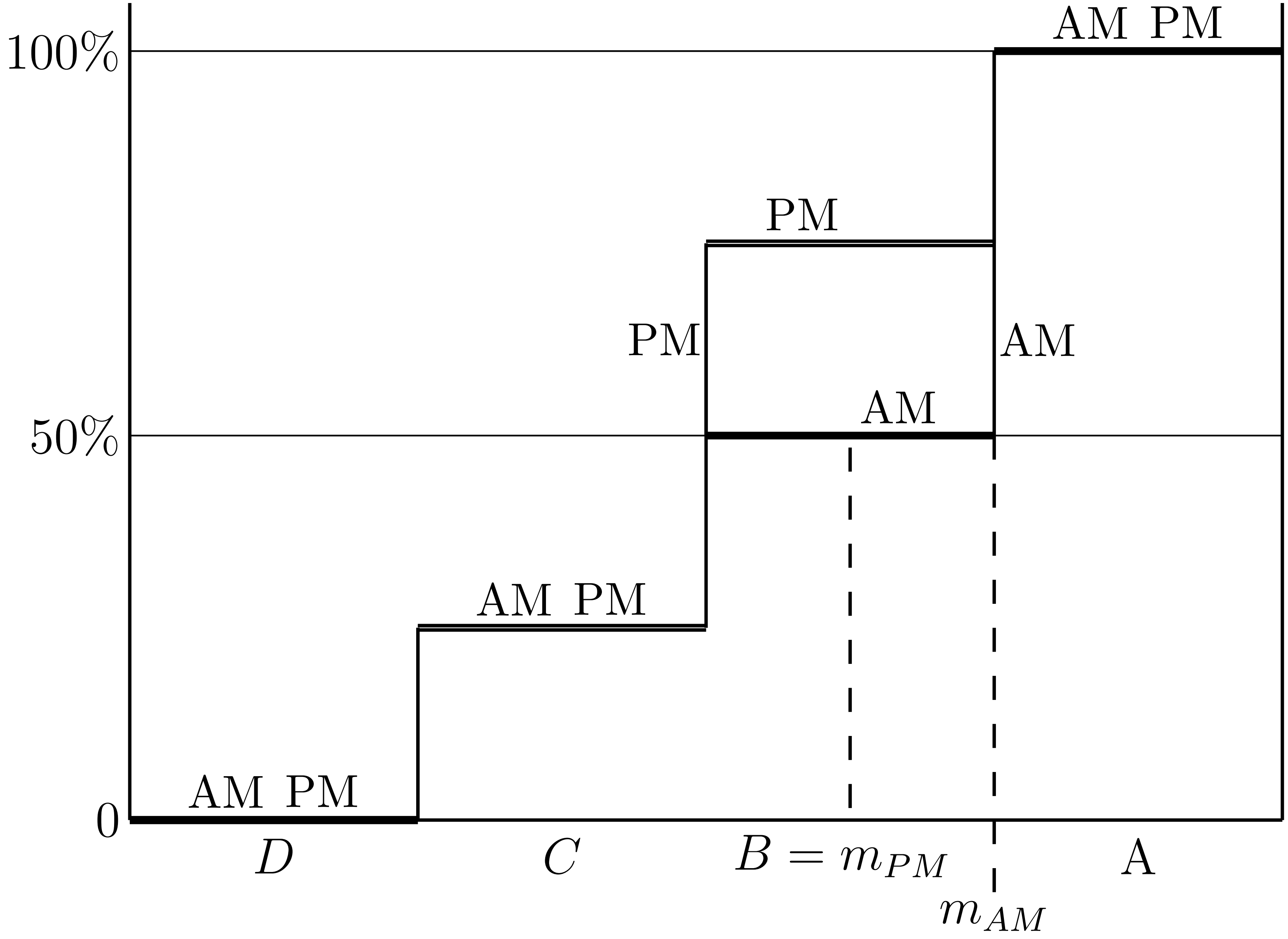
Grade	D	C	B	A
AM Class	0	2	4	8
PM Class	0	2	6	8

- and the relative CDFs:

Grade	D	C	B	A
AM Class	0.00	0.25	0.50	1.00
PM Class	0.00	0.25	0.75	1.00

- The median of the afternoon class is exactly $m_{PM} = B$, but that of the morning class m_{AM} is "between" A and B.

The CDF and medians, graphically



Merits and demerits of the median

- The median is *stable*; using different cells for the frequency counts will not affect the median.
- The median takes *all* values into account, not just the most frequent one.
- The median does not take the size of values into account. *E.g.* suppose you add 10 values whose size is 10 times the median, and 10 values at 99% of the median. The median does not move! (This can be an advantage or a disadvantage.)
- The median may be multivalued, but the median values are contiguous.

Percentiles: generalized median

- The median is the data value x that solves $F(x) = 1/2$, where F is the relative cumulative distribution function.

- There's nothing special about the rank "1/2," any other rank between 0 and 1 can be used.

- If the rank is a multiple of $1/4$, the corresponding data value is called a *quartile.* If the rank is a multiple of $1/10$, the data value is a *decile.* When the rank is expressed as a percentage, the corresponding data value is a *percentile.*

- In the equation $r = F(x)$, x is called a *percentile*, while r is the *percentile rank.* Usually it's obvious, but it can be confusing, especially with test scores and other data expressed in percentages.

Use of percentiles

- Percentile is often used as a synonym for cumulative distribution: “he is in the 80th percentile.” *I.e.*, 80% of people have lower scores than he. Strictly speaking this is incorrect. You should say: “*his score* is at the 80th percentile.” But this is rarely confusing.

- The median (50th percentile, second quartile) is a good choice for a *typical* value.

- Suppose you want to know whether America or Japan does a better job of providing for the poor. Then you could compare the 10th percentile (1st decile) of the two income distributions.
  - This is a measure of *location* of the two distributions; it is not a measure of “central tendency.”

The mean

- In mechanics, most problems can be solved by assuming that all mass in a body is concentrated at the *center of gravity*, regardless of the size, shape, or distribution of mass.

- In most applications of statistics, it is *incorrect* to make this kind of assumption, but it is often convenient and a good approximation.

- The *mean* of a distribution is the average of the values, each weighted by its frequency.

Computing the mean

- The mean is computed as a sum (or integral):

  $$\bar{x} = \frac{\displaystyle\sum_{x \in X} x f(x)}{\displaystyle\sum_{x \in X} f(x)}$$

  where X is the set of all values and f is the distribution function.

  - If f is the *relative* distribution function, then the denominator is identically 1.
  - This is convenient, so from now on all distributions used in calculating averages will be relative distributions.

- On a test, X may be all numbers $0 \leq x \leq 100$. In end-of-term reports, $X = \{A, B, C, D\}$, but we usually transform that to $X = \{4, 3, 2, 1\}$ precisely so we can compute the mean.

Measures of “spread” of a distribution

- We discussed the measures of location of distributions, including the use of CDF to compare locations of two distributions.
- We saw that the fact that the values in the distribution are “spread out” can lead to ambiguities or other undesirable behavior of our location measures.
- We claimed that the mode is a better measure of location when the distribution is *not* spread out.
- Measuring the “spread” or *dispersion* of a distribution is important.

Simple measures of dispersion

- The simplest measure of dispersion is the *range* of the distribution, which is the distance between the minimum and maximum observed values.

- The set of observed values is called the *support* of the distribution: $\mathrm{supp} f = \{x | f(x) \neq 0\}$.

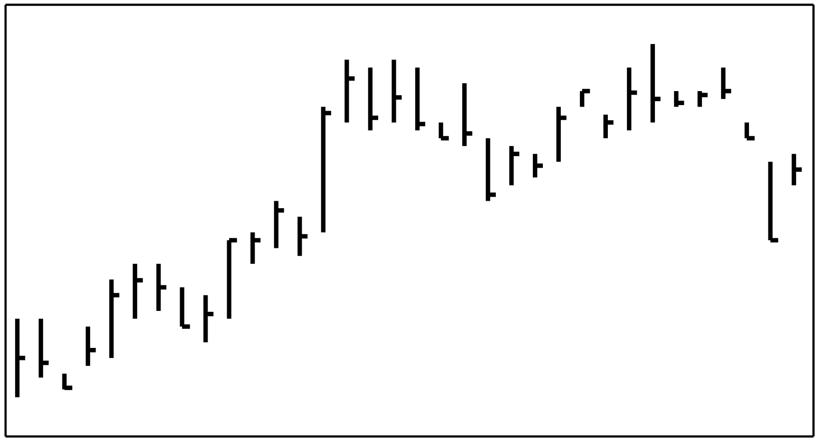
- The *range* is defined $r(f) = \max \mathrm{supp} f - \min \mathrm{supp} f$.

- Closely related to the range is the pair of extremes (min, max).

- Note that this meaning of range is different from the range of a function (the set the function values are taken from).

Example of the range

- The commonly used graph for the stock market can be interpreted as displaying the range and a measure of location. (What is the measure of location?)



Finer indicators of "spread"

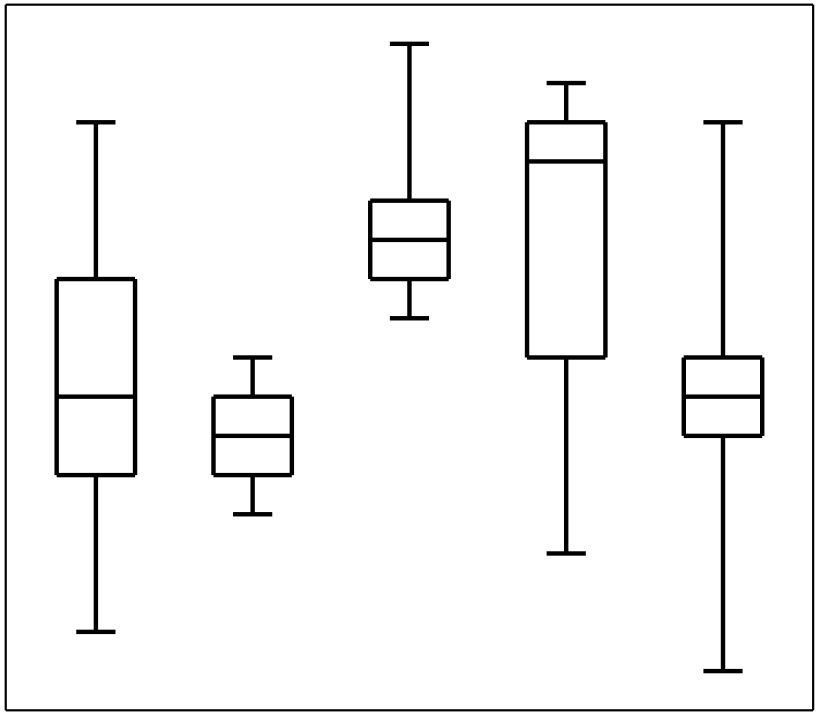
- From the point of view of statistics, the distribution tells us everything we need to know, everything we *can* know, about dispersion.

- The range has the problem that it is determined by two very unusual values, the extreme values. While this information is important, the extreme values are unstable, and the majority of values may be highly concentrated despite a large range. (In this situation we say the extreme values, and other values far from the main concentration, are *outliers*.)

Measure of dispersion based on quartiles

- One resolution is to add more information, but less than the whole distribution.
- Like the median, the *interquartile range* (the distance between the first and third quartiles) is stable in the sense of not being too influenced by a very small number of outliers.
- In particular, in a graphical display, use of the *quartiles* is very effective, for example a *box and lines graph.*

Example of box and line graph



- This kind of graph can be quite useful with time series or multivariate distributions (with another variable on the horizontal axis).

Using moments

- The range and the quartiles have the problem that they depend not only on the frequency of each value in the support, but also on comparisons of values (*i.e.*, *sorted* data). Sorting is both somewhat computationally intensive, and mathematically not "smooth" (calculus techniques don't work well).

- We'd like to use a *moment*, because moments are linear in the distribution.

- The first guess would be the "average difference" of the data from some representative value z. But this is just

$$\sum_{i=1}^{m}(x_i - z)f(x_i) = \sum_{i=1}^{m} x_i f(x_i) - \sum_{i=1}^{m} z f(x_i) = \bar{x} - z.$$

  This doesn't tell us how far the values are from each other.

Average distance doesn't work

- The next guess would be the "average distance" of the data from some representative value.

- We could use zero: $\sum_{x \in X} |x| f(x)$. It doesn't work well, since if all x are positive, this is just the mean.

- We could use the mean $\bar{x}$: $\sum_{x \in X} |x - \bar{x}| f(x)$. This is the *mean absolute deviation.* It can be used, but the non-differentiability at $\bar{x}$ (or zero) is inconvenient.

- It is especially difficult if we try to minimize the average distance for some reason (this is frequently the case in *estimation*), because the absolute value is linear except where nondifferentiable. Minimizing linear functions is computationally expensive.

Second moments

- It turns out that the right thing to do is to give observations more weight the farther they are from the mean. How much? Proportional to the distance, *i.e.*, the measure is the *mean squared deviation*, or *variance*:

$$s^2 = \sum_{x \in X} (x - \bar{x})^2 f(x).$$

- This is the *second central moment*, or the square of the (signed) distance from the "center" of the distribution, that is, the *mean*.
  - The first central moment is not useful, since it is identically zero: $\sum_{x \in X} (x - \bar{x}) f(x)$.
- Note that the variance is nonnegative by definition.

Standard deviation

- The variance is theoretically convenient for many purposes, but hard to give a practical interpretation.
  - Consider a distribution with equal numbers of observations at 50 and 150. Then the mean is 100. But the squared distance is $50^2 = 2500$ for *every* observation, so the variance is also 2500! What does that mean?
- If we take the square root of the variance, we get the *standard deviation*, in this case $s = 50$.
  - The standard deviation equals the mean absolute deviation (average distance), so it seems to be plausible to compare it to the mean as a distance.

More about the standard deviation

- An alternative formula for the standard deviation is

$$s = \sqrt{\sum_{x \in X} x^2 f(x) - \bar{x}^2}.$$

  (The part inside the square root is equal to the variance.)

- The standard deviation is easy to use both in theory and computation.

- It corresponds to our idea of *average* deviation pretty closely.

- The extra weight it places on large deviations turns out to be the "right thing" for statistical analyses.

The standard deviation and the mean

- There are many useful estimates that can be made with mean and standard deviation only.

- It is common to use the standard deviation as a unit of measure. For example, the *coefficient of variation* is $\bar{x}/s$. A data set or distribution converted to use the standard deviation as the unit of measure is said to be *standardized.* (*Hensachi* is a standardized variable.)

- Every *normal distribution* is completely characterized by its mean and standard deviation. (So are several others, but this is the most important case.)

More fun with moments

- Define the *n-th central moment* as

$$\mu_n = \sum_{x \in X} (x - \bar{x})^n f(x).$$

- We'll use these to define *skewness* (ν) and *kurtosis* (κ).

- Both *skewness* and *kurtosis* are primarily interesting (outside of specialized fields) because $\nu \neq 0$ or $\kappa \neq 3$ indicates a *non-normal distribution*.

- Many statistical computations are inaccurate for non-normal distributions, so checking these statistics is important.

Skewness

- The *skewness* of f is $\nu = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{s^3}$.
- Because of the cubes, asymmetric distributions (distributions where deviations below the mean are larger than above, or vice versa) will have non-zero ν, while symmetric distributions have $\nu = 0$. In particular, the *normal distribution* always has a skewness of exactly zero.
- Distributions with a lower bound (typically zero) are often skewed. The income distribution is skewed. Similarly for upper bounds.
  - What matters is how close the mean is to the bounds. Exams have bounds of 0 and 100. A *very hard* exam will have a distribution skewed to the right. A *very easy* exam will have a distribution skewed to the left.

Kurtosis

- The *kurtosis* of f is $\kappa = \frac{\mu_4}{\mu_2^2}$.
- Kurtosis is useful in characterizing the “flatness” or “peakedness” of a distribution.
- Kurtosis is always non-negative. The *normal distribution* always has a kurtosis of exactly 3.
- A high-peaked, fat-tailed distribution has $\kappa > 3$. A distribution with a lot of medium size deviation has $\kappa < 3$.
- Fat-tailed distributions are important in finance (they contribute to “excess volatility”), and in inventory management and marketing (they lead to market concentration, *e.g.*, Amazon.com).

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.