NOTES for Monday 2012.JAN.30
STAT-UB.0103

By the way, if the list $x_1, x_2, \ldots, x_n$ constitutes an entire population (rather than a sample) we define the standard deviation as $\sigma = \sqrt{\frac{1}{n}\sum(x_i - \mu)^2}$, where we use $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ as a replacement for $\bar{x}$. For now, do not be overly concerned with this notion. It's fairly rare to be dealing with a list that represents an entire population.

Here's an interesting question about the standard deviation. "Why *this* out of all possible kinds of calculations?" Indeed, other measures of variability have been proposed. For instance, you could do the mean absolute deviation. This is defined as $\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$. You could also do the MAD, or median absolute deviation. If $\tilde{x}$ represents the median, then the MAD is $\text{median}\{|x_1 - \tilde{x}|, |x_2 - \tilde{x}|, ..., |x_n - \tilde{x}|\}$. Some of these are available in Minitab.

The standard deviation however has many elegant and useful mathematical properties. Not the least is this wonderful useful fact (the empirical rule):

about $\frac{2}{3}$ of all the values in a list are within $\bar{x} \pm s$

about 95% of all the values in a list are within $\bar{x} \pm 2s$

Do we like big values of *s* or small values of *s*? It clearly depends on what you are doing.

If you are manufacturing an object, you would like its physical properties to have very small standard deviations. Thus, if you're making tin cans, you would like the diameters to have a very low standard deviation. If you're making electronic components, you would like their resistance values to have a very low standard deviation.

We like coins to have very small standard deviations for their thicknesses and for their diameters, as we use these in vending machines. As an odd curious fact, U.S. one-cent pieces have a large standard deviation of weights, as the composition was changed in 1982.

When should standard deviations be large? If you were forming a panel of consumers to taste orange juice, you would like the standard deviation of their ages to be large. If you want to explore the relationship between income and education, you would like the

standard deviation of the variable "years of education" to be large in your data set. We must admit that we also like large standard deviations in stock price changes.

One neat use of standard deviations is the computation of *z* scores. Suppose that $x_1, x_2, \ldots, x_n$ is a list of values with average $\bar{x}$ and standard deviation *s*. Then the *z* score corresponding to x_j is $\frac{x_j - \bar{x}}{s}$.

If the list 34.2, 30.9, 36.6, 42.0, …, 38.1 has $\bar{x} = 36.0$ and $s = 5$, then

the *z* score corresponding to 34.2 is $\frac{34.2 - 36.0}{5} = -0.36$

the *z* score corresponding to 30.9 is $\frac{30.9 - 36.0}{5} = -1.02$

the *z* score corresponding to 36.6 is $\frac{36.6 - 36.0}{5} = +0.12$

and so on

These can be very useful for comparing items with unequal standard deviations.

> Suppose that commodity *A* has prices with a daily mean of \$15 and an associated standard deviation of \$2. Similarly suppose that commodity *B* has prices with a daily mean of 90 cents and an associated standard deviation of 14 cents. If today *A* is selling for \$18.25 and *B* is selling for \$1.12, which is relatively more overpriced?
>
> This can be answered by examining the *z* scores. For *A*'s current price of \$18.25, the *z* score is $\frac{\$18.25 - \$15}{\$2} = 1.625$. For *B*'s current price of \$1.12, the *z* score is $\frac{\$1.12 - \$0.90}{\$0.14} \approx 1.5714$. Thus it would appear that *A* is relatively more overpriced (but not by much).

As contrasted to the empirical rule, we have an exact story in Chebyshev's inequality. In any list of values, we have

Proportion of values for which $| x_i - \bar{x} | \geq k\, s$

$$= \frac{\text{number of values of } x_i \text{ for which } \left| x_i - \bar{x} \right| \geq ks}{n} \leq \frac{1}{k^2}$$

This is only interesting for $k > 1$. So, using $k = 2$ this tells us that the proportion of values 2 standard deviations or more away from the mean cannot ever exceed $\frac{1}{4}$. This is a mathematical upper limit! In real data, the proportion is around 5%.

Chebyshev's theorem is much more useful for mathematical theorem-proving than it is for saying useful things about real numbers.

We need a word about outliers. By this term, we refer to data which don't really fit with the rest of the pattern. Such data may be anomalous and possibly destructive to whatever purposes we have in mind. ***We do not have an agreed-on definition for outliers.*** One possible operational way of dealing with this comes through the box plot. After all, the items denote by * or by the circles could be considered outliers. Some people might say that any value with a *z*-score below -3 or above +3 could be considered an outlier. What should be done with outliers?

Check them! At least make sure they correspond to correctly-processed data values. If values are in error, they must either be corrected or removed. If values cannot be corrected, they must be removed, but in such cases any report about the data should note these removals.

If outliers really are correctly coded, then there are two choices.

You can leave them in the analysis as is.
You could also remove them from the analysis. This can't be a secret. Removed points must get their own little stories.

Finally, let's warn that you should not be overly vigorous about removing data points. People who overdo the search for outliers tend to find them, often incorrectly.

In chapter 4, we will talk about probabilities. This is the basis for all work in statistics, and it certainly represents more difficult material. Let's begin by considering exactly what we mean by a probability. We might end up with the simple notion that probability is just a primitive concept that doesn't need too much in the way of explanation.

Probability can be a fairly deep philosophical idea. Indeed, Todhunter in the 19th century wrote a very long and exhausting treatise on this topic. We'll get around it just by noting that probability is a primitive concept.

An event is a simple descriptive sentence. Thus "the coin comes up heads" defines an event.

For any situation, we define the set of relevant events as the sample space. We'll use the symbol $\mathbb{S}$, but certainly Ω is common.

We still should note these:

If E is an event then $0 \le P(E) \le 1$.

We use $P(E) = 0$ for impossible events.

We use $P(E) = 1$ for guaranteed events.

We use E' for "not E" and then $P(E) + P(E') = 1$. Equivalently, $P(E') = 1 - P(E)$.

Other notations for E' are $\bar{E}$, E^c, and $\neg E$.

If E and F are events that cannot happen together, then $P(E \cup F) = P(E) + P(F)$.

See handout: Types of Probability (pamphlet on probability)

This discusses origin of probability and other simple ideas. You already have enough crude understanding of probability, so you may not need the deep formalism.

We can solicit an opinion, say about the coming Super Bowl. This will be the Boston Patriots against the New York Giants. Let's get some opinions.

Suppose that Patrick thinks that the Patriots will win, and he believes P(Patriots) = 0.65. He would be willing to make the bet $\begin{cases} \$65 \text{ on Patriots} \\ \$35 \text{ on Giants} \end{cases}$ with the winner getting the whole $100. If he's a gambler (and not necessarily a partisan of either team), he would be willing to take either side.

Now let's suppose that $\mathbb{P}$ (for Patrick) believes $\begin{cases}\$65 \text{ on Patriots} \\ \$35 \text{ on Giants}\end{cases}$ is the right bet, reflecting his opinion that P(Patriots will win) = 0.65. Suppose he meets $\mathbb{G}$ (Ginny), who happens to believe that P(Patriots will win) = 0.60, meaning that $\mathbb{G}$ believes that $\begin{cases}\$60 \text{ on Patriots} \\ \$40 \text{ on Giants}\end{cases}$ is the right bet. Let's suppose that neither really wants a fair bet; each wants an advantage. After haggling, the bet might end up as $\begin{cases}\$62 \text{ on Patriots } (\mathbb{P}) \\ \$38 \text{ on Giants } (\mathbb{G})\end{cases}$. In this bet, $\mathbb{P}$ will take \$62 on the Patriots; after all, he was willing to go as high as \$65 on the Patriots and is happy to get this bet for \$62. Similarly, $\mathbb{G}$ was willing to go as high as \$40 on the Giants and is happy to get this for \$38.

The final form of the bet will depend on bargaining ability.

We need to be able to manipulate a few probability concepts.

> *See handout*: Arbitrage of Probability (pamphlet on probability)

The essential idea is that people can disagree over probabilities, and this leads to betting!

Let's ask exactly what probability is used for in the business context. First of all, it's part of a basic technical language.

> In making decisions, we'd like to have numerical values. Thus, if we make an investment, we'd like to assess probabilities on the various possible outcomes. (If I'm considering the purchase of an index fund today, I'd like to base my decision on the probability that the DJIA will be at 13,000 in a month. If we're considering the opening of a toy store in a certain shopping center, we'd like to think about the probability that our competitor will open a store in a nearby shopping center.)
>
> Probabilities are also the ingredients of many decision-making models. For instance, we'll be talking about the binomial probability model, and this describes an enormous number of real-world situations.

You are likely familiar with the set theory ideas of intersection, union, and complement, $\cap$ and $\cup$ and $'$. It might help to note that the complementation symbol could appear in other forms; thus $E' = \overline{E} = E^c = \sim E = \neg E$. Venn diagrams can be helpful.

The SAMPLE SPACE is a complete listing of all the things that can happen (according to what we're interested in). In making such a list, we must cover all possibilities, and items must not "overlap." It helps if we can list the probabilities.

EXAMPLE: Flip coin. Sample space is S = {heads, tails}; also P(heads) = $\frac{1}{2}$ and P(tails) = $\frac{1}{2}$.

EXAMPLE: Flip two coins. There are several ways to write out the sample space. Try S_1 = { HH , HT , TH , TT } giving the results of coin_1 and coin_2 (whether distinguished or not); each event has probability $\frac{1}{4}$. Try S_2 = {0, 1, 2} in terms of the number of heads; P(0) = $\frac{1}{4}$, P(1) = $\frac{2}{4}$, and P(2) = $\frac{1}{4}$. Also could use S_3 = {same , diff} according to whether coins come down with same side showing or not; P(same) = $\frac{1}{2}$ and P(diff) = $\frac{1}{2}$. Most people like S_1 since it is the most elemental.

EXAMPLE: Roll two dice. Most people would use S_1 = { 1&1, 1&2, 1&3, … 6&6 }, giving the results of die_1 and die_2 (whether distinguished or not). Each of these outcomes has probability $\frac{1}{36}$. You could also use S_2 = { 2 , 3 , ..., 12 } based on the total of the two dice. Also S_3 = { same , diff } or S_4 = {odd total , even total} or S_5 = { even & even, even & odd, odd & even, odd & odd }. Here S_1 is most useful.

Here's the setup, meaning the sample space, for the two-dice problem:

	Die 2 ⚀	Die 2 ⚁	Die 2 ⚂	Die 2 ⚃	Die 2 ⚄	Die 2 ⚅
Die 1 ⚀						
Die 1 ⚁						
Die 1 ⚂						
Die 1 ⚃						
Die 1 ⚄						
Die 1 ⚅						

Each of these boxes has exactly the probability $\frac{1}{36}$.

By noting appropriate boxes, we can show P(total = 7) = $\frac{6}{36} = \frac{1}{6}$. Here's the picture:

	Die 2 ⚀	Die 2 ⚁	Die 2 ⚂	Die 2 ⚃	Die 2 ⚄	Die 2 ⚅
Die 1 ⚀						
Die 1 ⚁						
Die 1 ⚂						
Die 1 ⚃						
Die 1 ⚄						
Die 1 ⚅						

Six boxes are shaded, and the probability is correctly noted as $\frac{6}{36}$.

Similarly P(total = 4) = $\frac{3}{36} = \frac{1}{12}$. Here's the corresponding picture:

	Die 2 ⚀	Die 2 ⚁	Die 2 ⚂	Die 2 ⚃	Die 2 ⚄	Die 2 ⚅
Die 1 ⚀						
Die 1 ⚁						
Die 1 ⚂						
Die 1 ⚃						
Die 1 ⚄						
Die 1 ⚅						

Three boxes are shaded, and we've noted the probability as $\frac{3}{36}$.

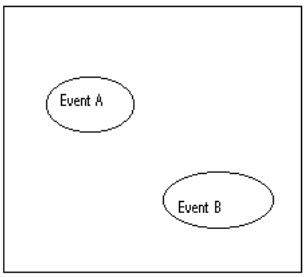
It's easy enough to enumerate *all* the possibilities for the total:

P(total = 2) = $\frac{1}{36}$ = P(total = 12)
P(total = 3) = $\frac{2}{36}$ = P(total = 11)
P(total = 4) = $\frac{3}{36}$ = P(total = 10)
P(total = 5) = $\frac{4}{36}$ = P(total = 9)
P(total = 6) = $\frac{5}{36}$ = P(total = 8)
P(total = 7) = $\frac{6}{36}$

You could make other versions of the sample space, but these other versions would not have the equally-likely property.

The use of MUTUALLY EXCLUSIVE EVENTS is helpful here. The definition is this: Events A and B are mutually exclusive if and only if $P(A \cap B) = 0$; that is, the events cannot happen together. This is sometimes described as *M.E.*

There is a simple picture, called a Venn diagram, which illustrates this. We show this by having A and B as non-overlapping regions.



In this picture, called a *Venn diagram*, the outer box represents the sample space S. The ovals inside represent events. The non-overlapping is used to suggest that $A \cap B$ has no probability.

If A and B are mutually exclusive, then $P(A \cap B) = 0$ (definition) and $P(A \cup B) = P(A) + P(B)$. This is a *special* form that only works for mutually exclusive events.

EXAMPLE: In the game of craps, one rolls two dice, with the interest in the total. You "crap out" on the first roll if your total is 2 or 3 or 12. Find

$$P(\{\text{total} = 2\} \cup \{\text{total} = 3\} \cup \{\text{total} = 12\})$$

These outcomes are M.E. (mutually exclusive), so you have

$$P(\{\text{total} = 2\} \cup \{\text{total} = 3\} \cup \{\text{total} = 12\})$$

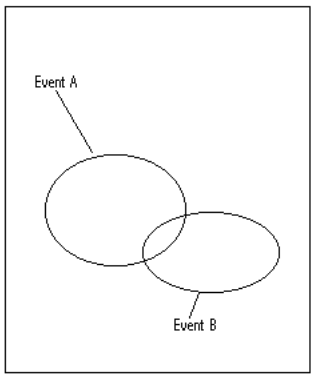
$$= P(\text{total} = 2) + P(\text{total} = 3) + P(\text{total} = 12)$$

$$= \tfrac{1}{36} + \tfrac{2}{36} + \tfrac{1}{36} = \tfrac{4}{36} \approx 0.11.$$

There is a general result which holds *always*:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This can be shown with this Venn diagram:



EXAMPLE: A standard deck of 52 cards has four suits (♠, ♥, ♦, ♣) with 13 cards in each suit. Each suit contains the ranks 2, 3, 4, …, 10, J, Q, K, A. The ranks J, Q, K are sometimes called "face cards." Suppose that you select one card at random and you are a winner if you get either a face card or a spade. That is, the event "win" is constructed as "face card" ∪ "spade." Then

$$P(\text{win}) = P(\text{face} \cup ♠) = P(\text{face}) + P(♠) - P(\text{face} \cap ♠)$$

$$= \frac{12}{52} + \frac{13}{52} - \frac{3}{52} = \frac{22}{52} = \frac{11}{26} \approx 0.4231$$

Consider a case in which there are 100 employees of a modest-sized firm. Let's suppose the cross-classified information is this:

	Firm's health plan	Spouse's health plan	TOTAL
Male	45	15	60
Female	23	17	40
TOTAL	68	32	100

Suppose that one employee is selected at random. Find P(male), P(female), P(firm's health plan), P(female or firm's health plan), …. These are all routine.