1. A special case of means: The proportion

A proportion could be thought of as the mean of a special kind of population. The population only has values of 1 or 0. If a population has that feature, the population mean is

p which is the proportion of 1's in your population.

And the population standard deviation is

$\sigma_p = \sqrt{p * q}$ where q is the value of (1.0 – proportion of p's)

For example:

```
    clinton |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        379       43.36       43.36
          1 |        495       56.64      100.00
------------+-----------------------------------
      Total |        874      100.00
```

```
                          clinton
-------------------------------------------------------------
      Percentiles      Smallest
 1%             0             0
 5%             0             0
10%             0             0       Obs                 874
25%             0             0       Sum of Wgt.         874

50%             1                     Mean           .5663616
                        Largest       Std. Dev.      .4958603
75%             1             1
90%             1             1       Variance       .2458775
95%             1             1       Skewness      -.2678155
99%             1             1       Kurtosis       1.071725
```

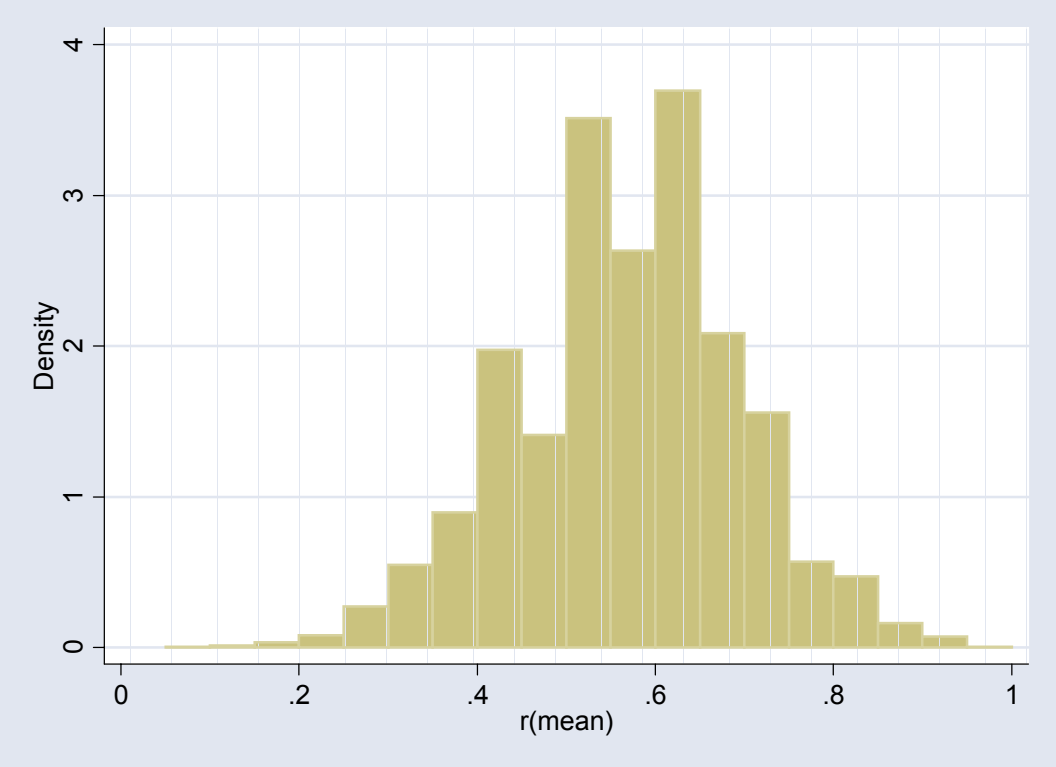
Proportions also have a sampling distribution, it's a distribution of sample proportions and this distribution has a mean of p and a standard deviation of

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

And if I were to run a simulation of samples of size 25 for 10000 samples

```
                          clinton
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .2666667       .0833333
 5%     .3529412       .0833333
10%           .4       .0909091       Obs               10000
25%     .4705882             .1       Sum of Wgt.       10000

50%     .5714286                     Mean           .5657526
                        Largest       Std. Dev.      .1289778
75%     .6470588       .9411765
90%     .7333333       .9444444       Variance       .0166353
95%     .7692308              1       Skewness      -.0970473
99%     .8571429              1       Kurtosis       2.948515
```



We can see that proportions behave like the mean, in theory it wants to center on the value of p (the true population proportion) and have a standard deviation of $\sigma_p = \sqrt{\frac{pq}{n}}$

2. The Margin of Error in the Press and the Confidence Interval in Statistics (p. 357)

"Margin of Error" as it is called in the media and "Confidence Interval" as it is called your book (p. 356-357) are closely related and they are both indicators of the "strength"/"believability"/"accuracy" of a statistic They are expressions of confidence in what conclusions you might draw from a survey result.

A "margin of error" as reported in the popular press is the same as a confidence interval with a confidence level of 95%.

3. Calculating the "Margin of Error" for a sample percentage

Look at the Bush poll, let's treat "vote for Bush" as a "success" (p) all other responses as a "failure" (q).

Let's then substitute our percentages from the sample ($\hat{p}$) for our population percentages p (which are actually unknown because we don't know what all voters will do tomorrow).

We can now calculate the standard error of $\hat{p}$ is $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{.48 * .52}{1003}} = .0158$

As long as n is large, and large is $n\pi \geq 10$ or $n(1-\pi) \geq 10$

Then the sampling distribution is approximately normal with mean p and standard deviation σ_p

Using our table of standard normal curve areas, we can find a value z* such that a CENTRAL area of .95 falls between +z* and –z* (and this means .025 in each tail area). The z* which would satisfy this is +1.96 and –1.96

For any normal distribution then, approximately 95% of the values are within +1.96 standard deviations of the mean. Since for large samples the sampling distribution of a sampling proportion $\hat{p}$ is approximately normal with a mean p

And $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ the confidence interval has the form $\hat{p} \pm Z^* SE(\hat{p})$

If we multiply the SE of .0158 by 1.96 (this is Z*) and then reported "48% + or – 3.1 percentage points" we have constructed a 95% confidence interval.

And what it says is that we expect 95% of all samples of this size to generate a statistic p that is within 3.1% of the true population proportion.

It implies that 95% of all possible samples of this size should generate statistics that fall within 3.1% of the "truth" (population parameter) AND that 5% of the time, the samples will generate statistics that fall outside of this interval.

The formula given in your text for a 95% confidence interval is: (p. 359)

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

this can be used as long as long as n is large, and large is $n\pi \geq 10$ or $n(1-\pi) \geq 10$. Note that when you substitute $\hat{p}$ for p , your book calls this "the standard error of a statistic" (p. 355) it's the estimated standard deviation of the sampling distribution.

Let's examine the other poll about the elderly's support of Bush (apparently 44% support him, the rest don't): We now calculate the standard deviation of the sampling distribution (actually standard error now because we're substituting) as

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{.44 * .56}{2002}} = .0111$$

And then multiplied the .0111 percent 1.96 and for that second survey, the result is reported "44% + or - 2 percentage points".

4. Confidence Interval Basics

A CONFIDENCE INTERVAL then is a range of values (i.e. values derived from sample information) that we think covers or contains the true parameter. The Washington Post says 48% of those surveyed would vote for Bush if the election were held tomorrow with a margin of error of 3%. This suggests a range around the sample statistic of 48% is 45% to 51%. This interval of 45% to 51% is supposed to

"cover" or "contain" the true percentage among American voters. This plus or minus 3% is effectively the same as plus or minus 2 Standard Errors and this is the way the media typically expresses results from polls. What they are saying is that they are "95% confident that the interval 45% to 51% covers or contains the true percentage of all voters who would cast a vote for Bush if the election were now.

In our other examples then, we are "95% confident that the interval 42% to 46% covers the true percentage of the elderly who support Bush" and we are "95% confident that the interval 43% to 57% covers the true percentage of Catholics who think the Pope should step down". All this means is the following: IF WE COULD REPEAT THIS PROCEDURE 100 TIMES, 95 OF THOSE TIMES (or 95%) OF THOSE TIMES, WE WOULD HAVE SELECTED A SAMPLE SUCH THAT THE CONFIDENCE INTERVAL GENERATED CONTAINS OR COVERS THE TRUE POPULATION PARAMETER. 5 OF THOSE TIMES, THE INTERVAL WE GIVE YOU DOES NOT! WE HOPE THAT THIS IS ONE OF THE GOOD 95 AND NOT ONE OF THE BAD 5.

The figures 48% plus or minus 3% (etc.) are confidence intervals for the population percentage and they are calculated from sample percentages and sample standard deviations.

The level of confidence (e.g. 68%, 90%, 95%, 99%) have corresponding Z- values (i.e. 1, 1.64 or 1.65, 1.96, 2.58)

5. Properties of Confidence Intervals (p. 360-361)

You can never been 100% confident. There is always the chance that you could have a very bad sample and are nowhere near the true population parameter.

Note that as confidence decreases, the interval grows shorter and it is less precise, as confidence level increases the interval grows wider and is more precise. As your sample size increases, your interval grows shorter and it is more precise. As your sample size decreases, the interval is wider and less precise. If your standard deviation increases, the interval becomes wider and less precise. As your standard deviation decreases, the interval becomes narrower and more precise.