Demo 30: Where Does That Root (p(1 – p)) Come From?

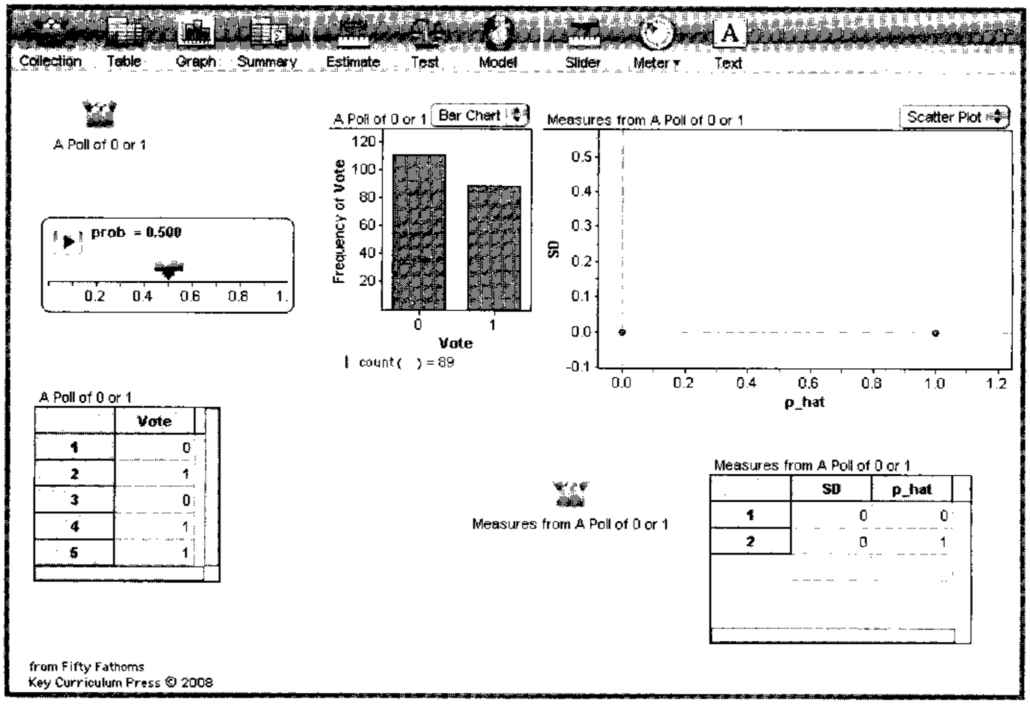
The standard deviation of a variable that's only 0 or 1 • Connecting the "proportion" situation to the "mean" situation

When you study the confidence interval for a proportion, you might learn this formula:

$$CI = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is the critical z-value (1.96 for a 95% confidence interval), $\hat{p}$ is the sample proportion, and n is the size of the sample. This formula assumes that n is large enough and that $\hat{p}$ is not too close to 0 or 1. (See Demo 31, "Why $np > 10$ Is a Good Rule of Thumb.")

If you ask the instructor what the $\sqrt{\hat{p}(1-\hat{p})}$ means, she may mutter, "It's the standard deviation." Yet it doesn't *look* like a standard deviation. If you press her, she may confess that "it is the standard deviation if you code a success as one and a failure as zero"—and leave it at that. What does that mean? That's what this demo is all about.



A Poll of 0 or 1

	Vote
1	0
2	1
3	0
4	1
5	1

Measures from A Poll of 0 or 1

	SD	p_hat
1	0	0
2	0	1

What To Do

▹ Open **SD of a Bernoulli Variable.ftm.** It will look something like the illustration.

Let's focus on the left half of the document first: It's a simple poll. The collection is 200 cases with one attribute: **Vote,** which is 0 or 1. The slider **prob** determines the chance that a **Vote** will be 1 instead of 0. You can see the results of the poll in the bar chart in the middle. Notice that even though **prob** is 0.500, the counts are not equal.

▹ Drag the slider and watch how the proportions of **Vote** change in the graph. (Change the bar chart to a ribbon chart if you think that's clearer.)

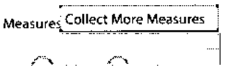
Note: Throughout this demo, we ignore thorny issues such as whether we should be using population SD or sample SD, whether we should use $\hat{p}$ or p, and so on. In this situation, those distinctions do not matter much. The point is that $\sqrt{p(1-p)}$ measures the spread in the set of 1's and 0's that make up the original data.

- Keep dragging the slider and also notice how the values for **Vote** change in the case table. Note that while you can see only five cases, there are 200 cases in all. (You can stretch the table to see more cases if you like.)

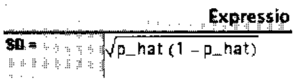
Note: **SD** is **popStdDev(vote)**; since we have 200 cases, this is very close to the sample SD.

Now let's look at the right half. The "measures" collection collects values for **p_hat**—the proportion of 1's—and the standard deviation of the sample, **SD,** whenever we tell it to. You can see that we have already collected two important data points: The standard deviation is 0 when **p_hat** is 0 or 1.

- Now let's tell the measures collection to collect more data. Set the slider **prob** to somewhere near 0.5.
- Click the measures collection once to select it.
- Drag the right edge of the measures collection to the right just a tiny amount. A **Collect More Measures** button will appear (as shown) as soon as the collection is no longer "iconified."



- Press the button. A new data point appears on the graph (lower right), and the new data appear in the table.
- Repeatedly move the slider and press the button to fill in the arc of points.
- When you're done adding points, put in the curve. First, click on the graph to select it.
- Choose **Plot Function** from the **Graph** menu. The formula editor appears.
- Enter $\sqrt{\textbf{p\_hat(1-p\_hat)}}$, as shown. Press **OK** to exit the editor. The curve appears and should go right through the points.



SD = √p_hat (1 – p_hat)

So, when you have just 1's and 0's, you can easily calculate the standard deviation by using that simple formula. The more profound lesson is that the formula for the CI of a proportion is really the same as the one for the mean: *The width of the confidence interval is* $t^*s/\sqrt{N}$. The formulas look different because of this shortcut for the standard deviation in the proportion case.

Questions

1 When you set **prob** to 0.500, the counts are not equal. Why not?

2 Why is the standard deviation 0 when **p_hat** is 0 or 1?

3 Does that mean that the confidence interval there has zero width? Why or why not? **Sol**

Challenges

4 You might wonder why we use 0 and 1 to represent the two states in a Bernoulli variable. Why not 0 and 2? Or −1 and +1? One reason is that then the *mean* of the data is conveniently equal to the proportion of 1's. Prove it.

5 We demonstrated that $\sqrt{\hat{p}(1-\hat{p})}$ was equal to the (population) standard deviation as computed by the computer. Let's check that analytically. You know that the standard deviation is really the square root of the variance, and that variance is the expected value of the squared deviation from the mean, that is,

$$\text{Var}[X] = E[(X-\mu)^2]$$

Show that this is equal to $p(1-p)$. **Sol**