In a Stern-Gerlach experiment, the arrival of an atom at a measurement counter is a random process. I would like to use the results of the experiments to answer the question:

What is the probability $\mathcal{P}$ that an atom will arrive at the top counter?

In the case where all the atoms arrive at the top counter, the probability is 1. However, what if I send 10 atoms through the analyzer and detect 3 atoms in the top counter? How confidently can I conclude that the probability is 0.3? What if I repeat my experiment and send 10 more atoms though the analyzer but detect 4 atoms in the top counter? I probably want to revise my estimate. If I do a bunch of sets of experiments, I will get a distribution of probabilities. Therefore, I'm going to need statistical tools to answer my questions:

1. What is the best estimate of the probability, given the experimental data?
2. How confident am I of that estimate?

To find the best estimate of the probability, I'm going to do a bunch of sets of experiments and take the mean. The mean probability will be my best estimate of the probability.

To determine how confident I am in the estimate, I'm going to consider the shape of the distribution. (For random processes like the Stern-Gerlach experiment - or coin flipping experiments, where there are 2 possible outcomes for each experiment - the underlying distribution is a *binomial distribution.*) To get a distribution, I can't do just one Stern-Gerlach experiment, or even a one set of Stern-Gerlach experiments - I have to do a bunch of *sets* of Stern-Gerlach experiments.

0.1 Some Definitions

$\mathcal{P}$ is the "true value" of probability of ending up in the top counter for the physical system (measuring $S_z = \hbar/2$). (This probability is the number that I'm trying to experimentally estimate.)

In 1 Stern-Gerlach experiment, as single atom passes through the analyzer and is detected at a counter.

I'm going to do a bunch of experiments and organize them into N sets. Each individual set n will include M particles being sent into an analyzer and counted in a counter.

For example, I can click the "10k" button and send 10,000 particles through the analyzer (i.e., 10,000 experiments). I can record the number of particles in the top counter and then repeat so that I end up with 5 sets of experiments.

Missing /var/www/paradigms_media_2/media/activity_media/spins_example_data.png

M = the number of Stern-Gerlach experiments in each set. This is the number of particles I send through the analyzer in 1 set. I'll assume that each set has the same number of experiments.

x_n = the (integer) number of atoms in the top counter after M Stern-Gerlach experiments

$\mathcal{P}_n$ is the probability I determine for 1 set of M Stern-Gerlach experiments.

N = the number of sets of Stern-Gerlach experiments (note: n is an index that indicates a single set of experiments)

$\bar{\mathcal{P}}$ is the mean probability determined from N sets of M experiments. This will be my estimation of the true probability.

0.2 Best Estimate of the Probability: the Mean

The probability I determine for a set of experiments (like in the table above) is:

$$\mathcal{P}_n = \frac{x_n}{M}$$

The mean of these probabilities is:

$$\bar{\mathcal{P}} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{P}_n$$

If I want to, I can also write the mean probability in terms of the number of atoms counted:

$$\begin{aligned}\bar{\mathcal{P}} &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{P}_n \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{x_n}{M} \\ &= \frac{1}{NM}\sum_{n=1}^{N}x_n \\ &= \frac{1}{M}\bar{x}\end{aligned}$$

0.3 Experimental Uncertainty - the Standard Error

In this section I'm going to argue that the standard error is a sensible thing to report as the experimental uncertainty (and for making statistical inferences). In order to understand the standard error, I'm first going to talk about the variance and the standard deviation.

0.3.1 The Variance

In order to quantify how spread out the distribution is, conceptually I'm tempted to find the average of the difference between each probability $\mathcal{P}$ and the mean of the distribution. The problem with this approach is that this average should be zero - the average is at the center of all the observations!

$$\frac{1}{N}\sum_{n=1}^{N}(\bar{\mathcal{P}}-\mathcal{P}_n) = \frac{1}{N}\left(\sum_{n=1}^{N}\bar{\mathcal{P}}\right) - \left(\frac{1}{N}\sum_{n=1}^{N}\mathcal{P}_n\right)$$

$$= \frac{1}{N}N\bar{\mathcal{P}} - \bar{\mathcal{P}}$$

$$= \bar{\mathcal{P}} - \bar{\mathcal{P}}$$

$$= 0$$

One way to get around this is to square all the differences first. The **variance** is the squared difference between the probability for one set of SG experiments and the mean probability:

$$var = \frac{1}{N}\sum_{n=1}^{N}(\bar{\mathcal{P}}-\mathcal{P}_n)^2$$

All contributions to the variance are positive, so the variance is greater than zero (though a zero variance is still technically possible if the distribution is one number). The larger the variance, the more spread out the distribution.

0.3.2 The Standard Deviation

The **standard deviation** is the square root of the variance.

$$SD = \sqrt{var}$$

$$= \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\bar{\mathcal{P}}-\mathcal{P}_n)^2}$$

$$\rightarrow \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\bar{\mathcal{P}}-\mathcal{P}_n)^2} \quad \text{for small N}$$

(For N $<$ 30ish, there are theoretical arguments about how the standard deviation of the sample underestimates the true standard deviation of the system, so the prefactor is front is made a smidge larger.)

- **The standard deviation does not decrease with more sets of experiments**. The standard deviation does not vary with N. The standard deviation comes from taking an average (you add up N things and then divide by N). As N increases, the standard deviation does not change with the number of experiments (it might fluctuate a little because of the random nature of additional experiments, especially if the total number of experiments is small, but if you plot SD vs. N (e.g., the number of particles in the top counter), the best fit line should have a near-zero slope). Therefore, the standard deviation is a characteristic of the system.

Missing /var/www/paradigms_media_2/media/activity_media/binomial_num_N.png

- **The standard deviation is a characteristic of the combined physical and measurement system**, including information about the distribution of the physical system and sources of random uncertainty during the measurement process.
- **Binomial vs "normal" distribution** For large numbers of experiments (M), a binomial distribution is very close to a normal (or Gaussian) distribution. For a normal distribution, 68% of measurements will lie within 1 standard deviation from the mean.

Missing /var/www/paradigms_media_2/media/activity_media/binomial_normal1.png

Missing /var/www/paradigms_media_2/media/activity_media/binomial_normal2.png

- **The standard deviation is a special kind of average, an *rms average*.** The rms stands for "root mean square" and describes the order of operations in the calculation (first you square, then you average, then you take a square root). So, the rms average allows me to get a sense of how far away individual probabilities $\mathcal{P}_n$ are from $\bar{\mathcal{P}}$ without running into the problem with doing a regular average, as described above.
- **Subtle difference between the distributions of number of atoms and probability.** The standard deviation *does* depend on the number of measurements in each set (which conceptually makes sense to me because the standard deviation is a characteristic of the combined physical and measurement system). For binomial distributions, the standard deviation for the distribution of the number of atoms is

$$SD_{x_n} = \sqrt{M\mathcal{P}(1-\mathcal{P})}$$

where $\mathcal{P}$ is the true probability I'm trying to measure. This equation for standard deviation is not general; it is only true for binomial distributions, where each experiment is a coin flip, atom through a Stern-Gerlach analyzer, etc. This equation tells me a system with a characteristic probability $\mathcal{P}$, the standard deviation will be twice as large if each set includes 100 experiments than if each set includes 25 experiments. (The mean will also be bigger because here I'm counting particles.)

Missing /var/www/paradigms_media_2/media/activity_media/binomial_num_M.png

In contrast, the standard deviation of the distribution of the **probabilities** is different by a factor of M

$$SD_{\mathcal{P}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\bar{\mathcal{P}}_n - \mathcal{P}_n)^2}$$

$$= \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{\bar{x_n}}{M} - \frac{x_n}{M}\right)^2}$$

$$= \sqrt{\frac{1}{M^2 N}\sum_{n=1}^{N}(\bar{\S_n} - \S_n)^2}$$

$$= \frac{1}{M}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\bar{x_n} - x_n)^2}$$

$$= \frac{1}{M}s_{x_n}$$

$$= \frac{1}{M}\sqrt{M\mathcal{P}(1-\mathcal{P})}$$

$$= \sqrt{\frac{\mathcal{P}(1-\mathcal{P})}{M}}$$

This tells me that the distribution of probabilities will get narrower as the number of experiments in each set gets larger.

Missing /var/www/paradigms_media_2/media/activity_media/binomial_prob_M.png

0.3.3 The Standard Error

The **standard error** (a.k.a. the standard deviation of the mean) σ is an a measure of how well I know the mean. In this lab, I'm estimating the true value of the probability by doing many (N) sets of Stern-Gerlach experiments and finding the mean of these sets. Now imagine that I repeat this whole process many times (N times) so that I get many means. Each mean is a better estimate of the true value of the probability than any individual probabily I measure, and the distribution of these means is much narrower than the distribution of the probabilities that I determined from each set of Stern-Gerlach experiments. If I compute the standard deviation of the distribution of means, it turns out that:

$$StErr_{\mathcal{P}} = \frac{SD_{\mathcal{P}}}{\sqrt{N}}$$

(see Taylor, pp. 147-148 for a nice derivation) If I only find one mean ($\bar{\mathcal{P}}$ from my original N sets of Stern-Gerlach experiments), I can be confident that there is a 68% chance that my mean lies is within

1 standard error from the true value of the probability.

In the case of Stern-Gerlach experiments (which following a binomial distribution):

$$StErr_{\mathcal{P}} = \frac{\sqrt{\frac{\mathcal{P}(1-\mathcal{P})}{M}}}{\sqrt{N}} = \sqrt{\frac{\mathcal{P}(1-\mathcal{P})}{MN}}$$

- **The standard error of the probability varies with the total number of experiments**. Notice that MN is the total number of Stern-Gerlach experiments that I run (M experiments in each set for N sets). The standard error is inversely proportionally to the square root of the total number of Stern-Gerlach experiments. It doesn't matter how I group them. If I do 10,000 Stern-Gerlach experiments, the standard error is the same as if I do 10 sets of 100 experiments, 20 sets of 50 experiments, or 10,000 sets of 1 experiment. It's hard to tell from looking at the plot alone that the standard error is the same:

  Missing /var/www/paradigms_media_2/media/activity_media/binomial_NMconstant.png

- **The standard error as a measure of uncertainty** The standard error tells me about how well my mean probability estimates the true value of the probability. Conceptually, it makes sense that the more experiments I do, the more confidence I should have in my estimate.

0.4 Reporting Uncertainty

To answer the question of how confident I am in my estimates, I *could* choose to report the uncertainty as the standard deviation or the standard error. These two options have different meanings (for this discussion, I'm going to assume that M is large and we have an approximately normal distribution):

$\bar{\mathcal{P}} \pm SD_{\mathcal{P}}$

Meaning: If I do one more set of SG experiments, there is a 68% chance that the probability I measure will fall in this range.

$\bar{\mathcal{P}} \pm StErr_{\mathcal{P}}$

Meaning: If I repeat the entire exercise, doing N sets of M SG experiments, there is a 68% change that the average probability I determine will fall in this range.

In this case, the standard error of the mean is closer to the thing I mean by my confidence in my estimate.

0.5 Comparing Values

If I wanted to compare my estimate of the probability to either (1) someone else's measurement or (2) a theoretically expected answer, both of which I'll call $\mathcal{P}_{exp}$, I might describe the difference between values in terms of the number of standard errors.

$$t = \frac{|\bar{\mathcal{P}} - \mathcal{P}_{exp}|}{StErr}$$

A smaller t corresponds to a higher likelihood that the two values come from the same normal distribution. The boundary between acceptable and unacceptable differences is a matter of opinion, to be decided by the experimenter (and the reader). For normal distributions, many scientists consider differences of:

$t < 2$ to be acceptable ("the discrepancy is insignificant") and

$t > 2$ to be unacceptable ("the discrepancy between values is significant.").

Differences that are $t \approx 2$ (1.9-2.6) are generally considered inconclusive.

For a normal distribution, there is a 95% likelihood that the true values lies with 2 standard errors of mean, meaning $t < 2$.

For Your Information:

Inferential statistical tests can be used to formally compare values, for example:

- **One Sample T-Test:** A one sample t-test allows us to test whether a sample mean (of a normally distributed variable) significantly differs from a hypothesized value.
- **Independent Samples T-Test:** An independent samples t-test is used when you want to compare the means of a normally distributed dependent variable for two independent groups.
- **Binomial Test:** A one sample binomial test can be used to determine whether the proportion of successes on a two-level categorical dependent variable significantly differs from a hypothesized value. (Remember that for large values of M, a binomial distribution approximates a normal distribution, so the first two tests might be applicable.)

For each statistical test, a set of assumptions need to be met in order for the test to give reliable, meaningful results. For example, a one sample t-test assumes that the data are normally distributed.