(1)

a. X is a binomial distribution with $n = 12000, p = 1/6$

b. The expected value, variance, and standard deviation of X is:

$$\begin{aligned}
\mathbb{E}(X) &= np = \frac{12000}{6} = 2000 \\
var(X) &= np(1-p) = 2000 \cdot \frac{5}{6} \approx 1666 \\
stdev(X) &= \sqrt{np(1-p)} = \sqrt{2000 \cdot \frac{5}{6}} \approx 41
\end{aligned}$$

c. Remember, a 95% confidence interval is a range of values for which we are 95% sure that the outcome of X (i.e. the number of sixes observed) will lie within. Since 95% of the distribution lines within two standard deviations of the mean, we get:

$$\begin{aligned}
Pr(2000 - 2 \cdot 41 \leq X \leq 2000 + 2 \cdot 41) &\approx 0.95 \\
Pr(2000 - 82 \leq X \leq 2000 + 82) &\approx 0.95
\end{aligned}$$

Again, this means that we are 95% sure that the number of sixes that will turn up out of 12000 tosses is somewhere between 1918 and 2082, or in other words, in the range of 2000 ± 82.

(2)

a. X is a binomial distribution with $n = 200, p = 0.01$.

b. The expected value, variance, and standard deviation of X is:

$$\begin{aligned}
\mathbb{E}(X) &= np = \frac{200}{0.01} = 2 \\
var(X) &= np(1-p) = 2 \cdot 0.99 \approx 1.98 \\
stdev(X) &= \sqrt{np(1-p)} = \sqrt{2 \cdot 0.99} \approx 1.41
\end{aligned}$$

c. Since with 99% probability, X lies between $\mathbb{E}(X) - 3 \cdot stdev(X)$ and $\mathbb{E}(X) + 3 \cdot stdev(X)$, then

$$Pr(2 - 3 \cdot 1.41 \leq X \leq 2 + 3 \cdot 1.41) \approx 0.99$$

Thus, we are 95% sure that the number of left-handed people, out of our sample of 200 is going to be between -2.23 and 6.23. Since it's impossible to have a negative number of people, we can express this range as around $[0, 6]$.

(3) From the information given, we can estimate p, the probability that a single coin flip turns up heads. Let K = number of heads, and $n = 10000$ tosses. Then a good estimate of p is simply $K/n = 5400/10000 \approx 0.54$.

Now, how sure are we that this is a biased coin, given that we tossed it 10000 times? To do this, we have to compute the confidence interval for the fraction of heads that *should* turn up if we have an *unbiased* coin. Let X be the number of heads that turns up for an unbiased coin after 10000 tosses. Then X follows the distribution $binomial(10000, 0.5)$. Thus, the mean, μ, is $10000 \cdot 0.5 = 5000$, and the standard deviation, σ, is $\sqrt{5000 \cdot 0.5} = 50$. With 95% confidence, we'd expect X to be within two standard deviations of the mean, or $X = 5000 \pm 100$. This means that the upperbound of the number of heads should be 5100.

Back to our scenario... we saw that out of 10000 tosses, 5400 of them were heads. This is quite a bit higher than the upperbound that we computed, 5100. Thus, with 95% confidence, we can say that the coin is biased (toward heads).

(4) Let K be the number of females our of a sample of n people. Then an estimate of the fraction of females is K/n. As shown in section 6.3 from lecture, we can compute the mean, variance, and standard deviation of K/n. After doing this computation, we see that with 99% confidence, the true fraction of people who are female is

$$\frac{K}{n} \pm 3\sqrt{\frac{p(1-p)}{n}}$$

Since we don't know what p is in the first place, we want to get rid of this from our computation. As shown in lecture, $p(1-p)$ can only be 1/4 at the most. Thus, we can rewrite our interval to take this into account:

$$\frac{K}{n} \pm 3\sqrt{\frac{p(1-p)}{n}} \quad = \quad \frac{K}{n} \pm 3\sqrt{\frac{1}{4n}} \quad = \quad \frac{K}{n} \pm \frac{3}{2\sqrt{n}}$$

The problem states that we want our range of estimation to be within 0.01 of the true number of females. This means that we want each side of the ± to be less than 0.01, or in other words, that $3\sigma < 0.01$. Then, we can solve for n:

$$\begin{aligned} \frac{3}{2\sqrt{n}} &\leq 0.01 \\ \sqrt{n} &\geq \frac{1}{0.01 \cdot (3/2)} \\ n &\geq \left(\frac{1}{0.01 \cdot (3/2)}\right)^2 \\ n &\geq 22500 \end{aligned}$$

Let's analyze our answer. We need to sample 22500 people such that the number of females we get gives us a fraction p that has a small interval of uncertainty (i.e. 1%) with a high confidence of 99%. This makes sense as the high constraint (high confidence and small range of uncertainty) requires an extremely large sample size.

(5)

a. The probability that some wedge does not receive a single dart is $\left(\frac{19}{20}\right)^{100}$. This saying that each of the 100 throws has a $\frac{19}{20}$ chance that it will miss wedge i.

b. Let's say $Y_j^i = 1$ if the j^{th} dart lands in wedge i and 0 otherwise. Then we have

$$\begin{aligned} \mathbb{E}(X_i) &= \sum_{j=1}^{100} \mathbb{E}(Y_j^i) = \sum_{j=1}^{100} Pr(Y_j^i) = \sum_{j=1}^{100} \frac{1}{20} = \frac{100}{20} = 5 \\ var(X_i) &= \sum_{j=1}^{100} var(Y_j^i) = \sum_{j=1}^{100} \left(\frac{1}{20} - \frac{1}{20}^2\right) = \sum_{j=1}^{100} \frac{19}{400} = \frac{19}{4} \\ stddev(X_i) &= \sqrt{var(X_i)} = \frac{\sqrt{19}}{2} \approx 2.18 \end{aligned}$$

c. With 95% confidence, the number of darts that fall in wedge i is going to be between $\mathbb{E}(X_i) \pm 2 \cdot stddev(X_i)$. To get an upperbound, we want the largest number within this range, which is simply $\mathbb{E}(X_i) + 2 \cdot stddev(X_i) = 5 + 2(2.18) \approx 9.36$. So, $X_i \leq 9$.

d. The expected value and variance of Y_i is as follows. Remember, Y_i is not a binary random variable in this case, so some of our previous shortcuts in finding mean and variance do not hold.

$$\mathbb{E}(Y_i) = 1 \cdot \frac{10}{20} + (-1) \cdot \frac{10}{20} = 0$$
$$\mathbb{E}(Y_i^2) = 1^2 \cdot \frac{10}{20} + (-1)^2 \cdot \frac{10}{20} = \frac{1}{2} + \frac{1}{2} = 1$$
$$var(Y_i) = \mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2 = 1 - 0 = 1$$

e. Recall that the distribution of i.i.d variables $X_1 + \cdots + X_n$ (where each X_i has a mean of μ and variance σ^2) is approximately $\mathcal{N}(n\mu, n\sigma^2)$. Similarly, $Z_r - Z_b$ can be written as the sum $Y_1 + \cdots + Y_{100}$, where $\mu = \mathbb{E}(Y_i) = 0$ and $\sigma^2 = var(Y_i) = 1$. Then $Z_r - Z_b \sim \mathcal{N}(0, 100)$.

f. Since $Z = |Z_r - Z_b| = |Y_1 + \cdots + Y_{100}|$, we can easily compute the mean, variance, and standard deviation of Z from the answers we've obtained from above:

$$\begin{aligned} \mathbb{E}(Z) &= |\mathbb{E}(Y_1) + \cdots + \mathbb{E}(Y_{100})| = 0 \\ var(Z) &= var(Y_1) + \cdots + var(Y_{100}) = 100 \\ stddev(Z) &= \sqrt{var(Z)} = 10 \end{aligned}$$

So, with 99% confidence, the value of Z is between $0 - 3(10) = -30$ and $0 + 3(10)$. However, since Z is non-negative, we take only the positive portion of this range and end up with $[0, 30]$.

(6) The probability that the target is hit exactly twice is:

$$\binom{10}{2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8$$

The first term chooses which two of the shots (out of ten) will be the ones to hit the target. The second and third terms give the probability that exactly two shots will hit the target (i.e. 2 shots hit the target and the other 8 shots do not hit the target).

(7) We can compute the complement probability: How many samples do we need in order for the probability of having no colorblind people to be less than 5%? To get this, we can solve for n below:

$$\begin{aligned} (1 - 0.01)^n &< 0.05 \\ \ln(0.99^n) &< \ln 0.05 \\ n \ln 0.99 &< \ln 0.05 \\ n &> \frac{\ln 0.99}{\ln 0.05} \\ n &> 298.3 \end{aligned}$$

So, n should be at least 299 people.

(8) The problem wants us to devise a test for two different things: whether or not John is guessing (call this H_0) and whether or not John has extrasensory powers (call this H_1). But obviously, if John does not have extrasensory powers, then he must be guessing. Thus, we want to set up our test such that the upper bound on our confidence interval for H_0 is equal to the lower bound on our confidence interval for H_1. Since $\frac{K}{n} = \frac{1}{2}$ for H_0, and $\frac{K}{n} = \frac{3}{4}$ for H_1, then we want to solve for n (the number of trials it

will take to find a confidence interval that satisfies this equation) below:

$$\begin{aligned}\frac{1}{2}+\frac{1}{\sqrt{n}} &= \frac{3}{4}-\frac{1}{\sqrt{n}}\\ \frac{2}{\sqrt{n}} &= \frac{1}{4}\\ n &= 64\end{aligned}$$

Now that we know how many times to sample, we can go back and figure out what the confidence interval is. For H_0, the confidence interval is in the range of:

$$\left[\frac{1}{2}-\frac{1}{\sqrt{64}},\frac{1}{2}+\frac{1}{\sqrt{64}},\right]=\left[\frac{3}{8},\frac{5}{8}\right]$$

and the confidence interval for H_1 is in the range of:

$$\left[\frac{3}{4}-\frac{1}{\sqrt{64}},\frac{3}{4}+\frac{1}{\sqrt{64}},\right]=\left[\frac{5}{8},\frac{7}{8}\right]$$

Just as we had set it up, the upper bound of H_0 is equal to the lower bound of H_1, which is $5/8$. This means that we ask John to guess on $n = 64$ cards; if he guesses more than $5/8$ cards correctly, then we believe he has extrasensory powers since this number is within the confidence interval of H_1. If he guesses less than this, then with 95% confidence he is guessing, since it falls within the confidence interval of H_0.

(9)

a. Let's say that N is an estimate for the true number of fish in the lake. Then the fraction of fish with red spots should be $1000/N$. On our second draw, we see that $Z/100$ fish have red spots. The fraction of fish with red spots should, in theory, be similar, so we can set these equal to each other, and solve for N:

$$\begin{aligned}\frac{1000}{N} &= \frac{Z}{100}\\ N &= \frac{100000}{Z}\end{aligned}$$

b. Our variable Z should represent that out of 100 samples, we have a probability of $1000/N$ getting a fish with a red spot. Thus, the binomial parameters are $n = 100, p = 1000/N$.

c. If we can give a confidence interval for the true fraction of fish with red spots, then we can also give a confidence interval for the true number of fish. Since we obviously don't know the true sample size, N, we sampled a set of $n = 100$, and then found the number of fish with red spots out of that sample set, Z. This tells us that the fraction of fish with red spots should be something around $Z/100$ – but since this is just an estimate, we still need to find a confidence interval for it.

As shown in lecture (section 6.3), after computing the expectation, variance, and standard deviation of Z/n, we can assert with 95% confidence that the true fraction of fish should be in the range of:

$$\frac{Z}{n}\pm\frac{1}{\sqrt{n}} \quad = \quad \frac{Z}{100}\pm\frac{1}{10} \quad = \quad \frac{Z\pm 10}{100}$$

This fraction should be equal to the total number of fish with red spots (which is 1000) out of the total number of fish in the lake (which is N, what we are trying to solve for). Setting these fractions equal and solving for N, we get:

$$\begin{aligned}
\frac{Z \pm 10}{100} &= \frac{1000}{N} \\
N &= \frac{100000}{Z \pm 10}
\end{aligned}$$

(10) Let $K =$ the number of correct runs. Then, $K \sim \mathit{binomial}(n, 2/3)$, and $\mathbb{E}(K) = 2/3n$. This means that most of the time, returning the majority answer will give the correct answer since $(2/3)n$ correct runs counts as majority. However, we want to make sure that the probability of error (when the number of correct runs is less than $(1/2)n$, i.e. the majority) is less than 5%. So, we use a 95% confidence interval – we want to be 95% sure that $K > (1/2)n$. Since we know that with 95% confidence, K is between the range of $np \pm 2\sqrt{np(1-p)}$, we want to set the lower bound of this range to be greater than 1/2, and then solve for n:

$$\begin{aligned}
np - 2\sqrt{np(1-p)} &= \frac{1}{2}n \\
\frac{2}{3}n - 2\sqrt{n \cdot \frac{2}{9}} &= \frac{1}{2}n \\
2\sqrt{n \cdot \frac{2}{9}} &= \frac{1}{6}n \\
n \cdot \frac{2}{9} &= \left(\frac{n}{12}\right)^2 \\
n &= \frac{2 \cdot 12^2}{9} \\
n &= 32
\end{aligned}$$

This means that if we run $\mathcal{A}(x)$ 32 times, there is only 5% chance that the correct answer will come up less than $32/2 = 16$ times, making the algorithm give a final incorrect answer.