1 (a). The choice was not to include a uniform height scale between different categories - each category has its own scale, so the bars in different categories cannot be compared to one another visually.

1 (b). The following line of code will produce a bar chart with a common horizontal scale between all categories, allowing comparison between data from different categories.

```
daily.barh('Year')
```

2. For the version where we are drawing 3 samples from a set of 4 colors:

There are 4 different ways to get all the same color on 3 samples:

$$P(\textit{all same color}) = P(\textit{all blue}) + P(\textit{all green}) + P(\textit{all red}) + P(\textit{all purple})$$

Now, let's consider one of these probabilities. In order to get 3 blue's, we have to get a blue on the first draw, a blue on the second draw, and a blue on the third draw. Since we're sampling with replacement, each of these happens with probability ¼:

$$P(\textit{all blue}) = P(\textit{first draw blue}) * P(\textit{second draw blue}) * P(\textit{third draw blue}) = ¼ * ¼ * ¼$$

We multiply here because we are taking proportions of possible outcomes: Out of all possible outcomes, ¼ of them draw blue on the first draw; of all outcomes that have a blue on the first draw, ¼ of them have a blue on the second draw, etc.

The same is true for all four colors, so we get

$$P(\textit{all same color}) = ¼^3 + ¼^3 + ¼^3 + ¼^3 = 4 * ¼^3 = \frac{1}{8}$$

Alternate explanation: The chance that all the draws are the same color is the same as the chance that every draw matches the first draw:

$$P(\textit{all same color}) = P(\textit{draw } 1 = \textit{draw } 1) * P(\textit{draw } 2 = \textit{draw } 1) * P(\textit{draw } 3 = \textit{draw } 1)$$

The chance that the first draw matches itself is always 1. The chance that the second draw matches the first is the chance that the second draw is ¼, since there's only 1 color out of 4 that matches the first draw, and the third draw is the same.

$$P(\textit{all same color}) = 1 * ¼ * ¼ = \frac{1}{8}$$

For other versions of the exam, the same reasoning applies. In each case, the solution is:

$$\left(\frac{1}{\textit{<num colors>}}\right)^{\textit{<num samples>} - 1}$$

3. Important criteria for a statistic to be a good estimate of a population ***parameter*** include ***low*** bias and low ***variability***. The ***empirical*** distribution of a large ***random sample*** is likely to resemble the ***distribution*** of the population.

4 (a). Bin [4, 12) contains more players because the corresponding bar has a larger area than that of bin [2, 4).

The area of the first bar is less than 24 (2 times something less than 12) and the area of the second is greater than 24 (8 times something more than 3).

4 (b) (i). For the version of the midterm where the bins are [4, 6) and [6, 12):

$$((56/417) * 100) / 2$$

4 (b) (ii). For the version of the midterm where the bins are [4, 6) and [6, 12):

$$A = (2-0)*(17.63) + (4-2)*(11.39) + (6-4)*(\textit{part i}) + (18-12)*(1.6) + (26-18)*(0.45)$$
$$(100 - A) / (12-6)$$

Alternatively,

$$((12-4)*(3.6) - (6-4)*(\textit{part i})) / (12-6)$$

5. **Of course not! They should quit.**
Plausible explanation: Those who have quit had a good reason to (it's very difficult to quit even if you know you need to): their health became so bad from smoking, that this habit was impeding w/ everyday activity. They are on the path to recovery, but are coming from an even more adverse condition.

6 (a). **Null hypothesis:** The method is not biased. Appointments are equally likely to be assigned to the new system as they are to the old; i.e, there is an equally likely chance that an appointment is made on Monday, Wednesday or Friday as there is on Tuesday, Thursday, Saturday. (Note: the null hypothesis must be a fully specified chance model)
Alternative hypothesis: The method is biased. Appointments are *not* equally likely to be assigned to the new system as they are to the old; i.e., the method is biased against appointments made on Monday, Wednesday or Friday. (Note: the alternative specifies the direction of the bias, in this case against appointments made on Monday, Wednesday or Friday)

6 (b).

```
values = make_array()
for i in np.arange(10000):
    sample = np.random.choice(array2, 477)
    number_new = np.count_nonzero(sample == "New")
    values = np.append(values, number_new)
empirical_P = np.count_nonzero(values <= 206)/10000
empirical P
```

For other versions of the exam:
477 corresponds to 492 (white), 489 (off-white)
sample corresponds to chosen (white), draws (off-white)
10000 corresponds to 15000 (white), 5000 (off-white)
206 corresponds to 211 (white), 212 (off-white)

7. “Based on probability (+1 = ⅓, -1 = ⅓, 0 = ⅓), the net gain should be zero (not [the histogram centered at 20]). The further we deviate from zero, the less the probability should be (probability of winning $60 is $(⅓)^{60}$, which is very small, so not [the histogram that’s roughly uniform on -$60 to $60]). There is a chance of winning -$60 to $60, so the range should be greater than that of [the histogram that went from -$1 to $1]. This leaves [the histogram that’s bell-shaped and centered at 0], which makes sense, as it is central at 0, follows a Gaussian distribution, and has a reasonable range (chance of winning > $20 and < -$20 is too low to see).”

8.

```
where('State', are.equal_to('NY')).select('Opinion')
where('State', are.equal_to('CA')).select('Choice')
where('State', are.equal_to('TX')).select('Preference')

group('Opinion').column('count') / voters_NY.num_rows
group('Choice').column('count') / ca_voters.num_rows
group('Preference').column('count') / voters_tx.num_rows

0.5 * np.sum(np.abs(dist_CA - dist_NY))
0.5 * np.sum(np.abs(fl_dist - ca_dist))
0.5 * np.sum(np.abs(dist_ca - dist_tx))
```