Due: Tuesday, December 15, 1992

1. (8 points) One way to estimate the size of a set $X = \{x_1, ..., x_n\}$ is to sample the elements of $X$ from a uniform distribution with replacement until an element is sampled twice. The number of elements drawn will probably be larger for larger $n$.

$$k \leftarrow 0$$
$$S \leftarrow \emptyset$$
$$a \leftarrow \text{Uniform}\{X\}$$

**repeat**
$$k \leftarrow k + 1$$
$$S \leftarrow S \cup \{a\}$$
$$a \leftarrow \text{Uniform}\{X\}$$

**until** $a \in S$

We can consider $k$ to be a random variable. (By the way, the expected value of $k$ tends to $\frac{\pi}{2}\sqrt{n}$, so a predictor for $n$ given $k$ is $\frac{2k^2}{\pi}$).

A) What is $Pr\{k > 1\}$?

B) For arbitrary $j$, $1 \leq j \leq n$,

- In how many ways can we make $j$ distinct choices for the first $j$ elements?
- In how many ways can we make $j$ (not necessarily distinct) choices for the first $j$ elements?
- What is $Pr\{k \geq j\}$?

C) For $n = 10$, give numerical values for each of $Pr\{k = 1\}$, $Pr\{k = 2\}$, ..., $Pr\{k = 10\}$.

D) Write a program to estimate these ten probabilities by simulating the above experiment a (reasonably large) number of times.

E) Perform a $\chi^2$ test to determine whether your experiment provided a reasonable approximation to the ten probabilities.

2. (2 points) 5% of the disk controllers produced by a plant are known to be defective. A sample of fifteen controllers is drawn randomly from each month’s production and the number of defectives is noted. What proportion of these monthly samples would be expected to have at least two defective controllers?
3. (5 points) (From Cormen, Leiserson & Rivest) With a $t$-bit counter, we can normally count up to $2^t-1$. With a probabilistic counter, we can go much higher, with a loss of accuracy and precision. When a new counter $\kappa$ has the value $k$, this will denote a count of $n_k$ of $2^k-1$, for $k=0,1,...,2^t-1$. We let $\kappa$'s initial value be 0, corresponding to a count of $n_0=0$. To perform an INCREMENT of $\kappa$:

- **If** $\kappa$ contains $2^{t-1}$ **then** report overflow
- **else** increment $\kappa$ by 1 with probability $\frac{1}{n_{k+1} - n_k}$

If we choose $n_k=k$, then the counter is an ordinary one. More interesting situations arise if we choose things like $n_k=2^k-1$ for $k>0$. For the sequel, assume that $n_{2^t-1}$ is large enough that the probability of overflow is negligible.

Assuming that $n$ is one of the values $n_k$, show that the expected value of $n_k$ after $n$ INCREMENTs have been performed is $n$. 

1. A) \(\frac{n-1}{n}\)

B) \(\frac{(n-j)!}{n^j}\)

C,D) Using ThinkC for \(N=10,000\) trials:

<table>
<thead>
<tr>
<th>(j)</th>
<th># Samples</th>
<th>(Pr{k=j})</th>
<th>(N\times Pr{k=j})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>962</td>
<td>0.1</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>1804</td>
<td>0.18</td>
<td>1800</td>
</tr>
<tr>
<td>3</td>
<td>2154</td>
<td>0.216</td>
<td>2160</td>
</tr>
<tr>
<td>4</td>
<td>2022</td>
<td>0.202</td>
<td>2016</td>
</tr>
<tr>
<td>5</td>
<td>1527</td>
<td>0.151</td>
<td>1512</td>
</tr>
<tr>
<td>6</td>
<td>943</td>
<td>0.091</td>
<td>907</td>
</tr>
<tr>
<td>7</td>
<td>417</td>
<td>0.042</td>
<td>423</td>
</tr>
<tr>
<td>8</td>
<td>136</td>
<td>0.015</td>
<td>145</td>
</tr>
<tr>
<td>9</td>
<td>33</td>
<td>0.003</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.000363</td>
<td>4</td>
</tr>
</tbody>
</table>

\(\chi^2 = 4.456\) The fit for 9 degrees of freedom should be worse than this >80% of the time.

2. Given that the probability that a controller is not defective = 0.95, we want the probability that 14 or 15 of 15 independently chosen controllers are not defective. This is

\[1 - \sum_{14 \leq k \leq 15} \binom{n}{k} p^k q^{n-k} = 1 - \sum_{14 \leq k \leq 15} \binom{n}{k} (0.95)^k (0.05)^{n-k} = 0.17096\]

3. Define a hit to be the event that \(k\) gets incremented by 1, and let the \(k^{th}\) stage consist of the INCREMENTs between the \((k-1)^{st}\) hit and the \(k^{th}\) hit. In the \(k^{th}\) stage, \(Pr\{\text{hit}\} = \frac{1}{n_k - n_{k-1}}\). Let \(t_k\) be a r.v. denoting the number of INCREMENTs in the \(k^{th}\) stage. \(t_k\) is geometrically distributed, with probability of success \(\frac{1}{n_k - n_{k-1}}\).
\[ E[t_k] = \frac{1}{n_k - n_{k-1}} = n_k - n_{k-1}. \]

Defining r.v. \( X_p = \sum_{1 \leq m \leq p} t_m \), we see that \( X_p \) is a r.v. denoting the number of INCREMENTS of the first \( p \) stages.

\[
E[X_p] = E\left[\sum_{1 \leq m \leq p} t_m\right] = \sum_{1 \leq m \leq p} E(t_m) = \sum_{1 \leq m \leq p} (n_m - n_{m-1}) = n_p - n_0 = n_p
\]