Lecture 8

Sampling with replacement

Sampling without replacement

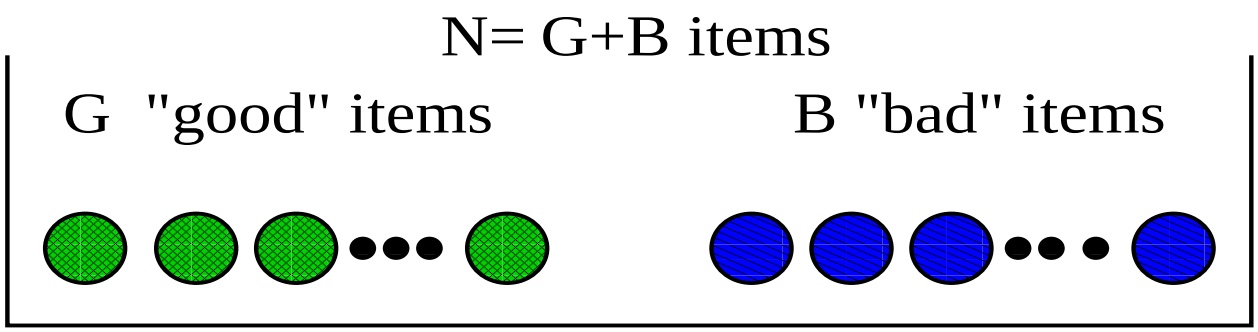
Hypergeometric distribution

Binomial approximation

Sampling many types

Random variables

Sampling



- If we choose n items out of the above box, with replacement what is the probability that we choose g "good" ones (and $b = n - g$ "bad" ones)?

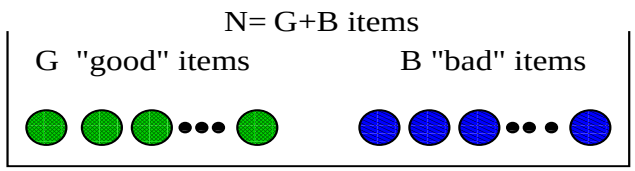
○ $0 \leq g, b \leq n < \infty$.

- If we choose n items out of the above box, without replacement what is the probability that we choose g "good" ones (and $b = n - g$ "bad" ones)?

○ $0 \leq g \leq \min(n, G), \quad 0 \leq b \leq \min(n, B),$

. $\quad 0 \leq n \leq N = G + B.$

Sampling with replacement



Solution 1: A binomial approach.

- Each drawing is a "success/failure" event, with the understanding "success" $\equiv$ "good" and "failure" $\equiv$ "bad".

$$\Rightarrow \quad p = \frac{G}{G+B} = \frac{G}{N}, \quad q = 1 - p = \frac{B}{G+B} = \frac{B}{N}.$$

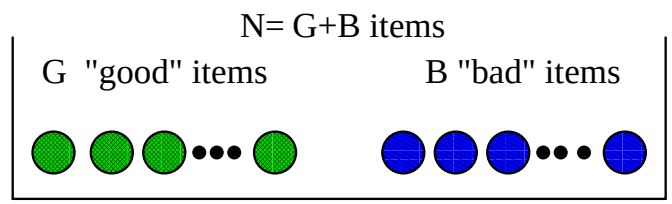
$$\begin{aligned} \Rightarrow P\{g \text{ good and } b \text{ bad}\} &= \binom{n}{g}\left(\frac{G}{N}\right)^g \left(\frac{B}{N}\right)^{n-g} \\ &= \binom{n}{g}\frac{G^g B^{n-g}}{N^n} = \binom{n}{g}\frac{G^g B^b}{N^n} \end{aligned}$$

Solution 2: A counting approach.

- We have to choose g good items and b bad items in n trials with replacement.
- Take Ω to be the space of all N^n possible choices of items. Since <u>each item</u> has the same probability of being chosen at any trial, $(\Omega, \mathcal{F}, \mathcal{P})$ is a <u>uniform</u> probability space.
- One each drawing, there are G ways to choose a good item, and B ways to choose a bad.
- There are $\binom{n}{g}$ ways to decide in which of the n draws we will see g good and b bad items.

$$\Rightarrow \qquad P\{g \text{ good and } b \text{ bad}\} \quad = \quad \frac{\binom{n}{g} G^g B^b}{N^n}$$

Sampling without replacement (hypergeometric)



Solution 1: A probability approach.

- Probability of sequence $\underbrace{GG\ldots G}_{g \text{ times}}\ \underbrace{BB\ldots B}_{b \text{ times}}$ is

$$\frac{G}{N}\frac{G-1}{N-1}\cdots\frac{G-g+1}{N-g+1}\frac{B}{N-g}\frac{B-1}{N-g-1}\cdots\frac{B-b+1}{N-g-b+1} \equiv \frac{(G)_g(B)_b}{(N)_n}$$

- Note: Probabiility is <u>independent</u> of order.
- There are $\binom{n}{g}$ possible orderings.

$$\Rightarrow \quad P\{g \text{ good and } b \text{ bad}\} \quad = \quad \frac{\binom{n}{g}(G)_g(B)_b}{(N)_n} \equiv \frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}}$$

$$0 \leq g, b \leq n, \quad g + b = n \leq N, \; (M)_m \equiv M(M-1)\cdots(M-m+1).$$

Solution 2: A counting approach.

- We have to choose g good items and b bad items in n trials without replacement.
- Take Ω to be the space of all $\binom{N}{n}$ possible sequences of items. Since, <u>from the point of view of sampling</u> all items are indistguishable, $(\Omega, \mathcal{F}, \mathcal{P})$ is a <u>uniform</u> probability space.
- There are $\binom{G}{g}$ ways to choose the good items.
- There are $\binom{B}{b}$ ways to choose the bad items.

$$\Rightarrow \qquad P\{g \text{ good and } b \text{ bad}\} \quad = \quad \frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}}$$

$0 \leq g, b \leq n, \quad g + b = n \leq N.$

Example: Capture-recapture

$$
\begin{aligned}
N &= \text{Unknown number of fish in the lake.} \\
n_1 &= \text{Number of fish in first catch.} \\
n_2 &= \text{Number of fish in second catch.} \\
t &= \text{Number of tagged fish in second catch.}
\end{aligned}
$$

$$P\{t \text{ tagged fish in second catch}\} = \frac{\binom{n_1}{t}\binom{N-n_1}{n_2-t}}{\binom{N}{n_2}}$$

- Suppose $n_1 = n_2 = 1000$ and $t = 100$. Then $N \geq 1900$, but taking $N = 1900$ implies

$$P\{100 \text{ tagged in 2nd catch}\} = \frac{\binom{1000}{100}\binom{900}{900}}{\binom{1900}{1000}} = \frac{(1000!)^2}{100!\,1900!} \approx 10^{-430}$$

$\Rightarrow \quad N = 1900$ is unreasonable.

Finding N

- Set $P_N = P\{100 \text{ tagged in 2nd catch}\} = \frac{\binom{1000}{100}\binom{N-1000}{900}}{\binom{N}{1000}}$.
- Maximise this as a function of N (Maximum likelihood) to find $N \approx 10,000$.

Stirling's formula

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

$$\begin{aligned} 100! &\sim \sqrt{2\pi 100}\left(\frac{100}{e}\right)^{100} \\ &= 25.067(36.787)^{100} = 25.067(10^{1.567})^{100} \approx 2.5 \times 10^{158} \end{aligned}$$

Feller: An Introduction to Probability Theory and Its Applications

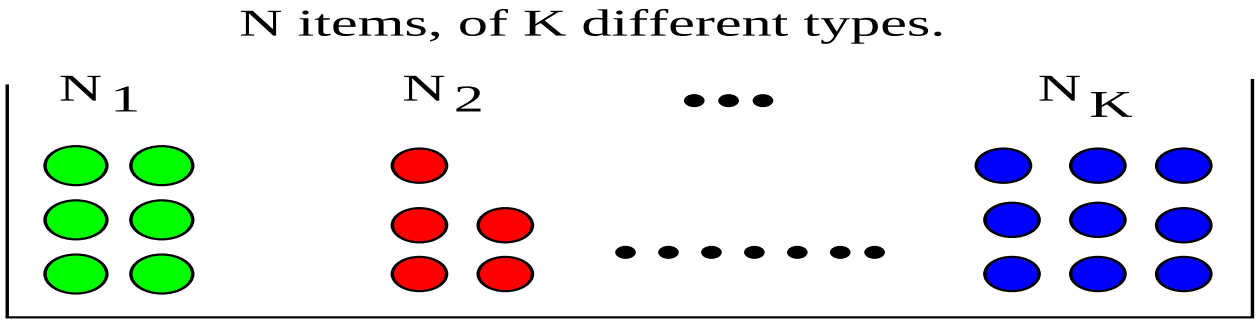
Binomial approximation

- We are in the sampling <u>without</u> replacement case.
- We assume that G and B are large with respect to g and b.

$$
\begin{aligned}
\frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}} &= \frac{G!}{(G-g)!g!}\,\frac{B!}{(B-b)!b!}\,\frac{(N-n)!n!}{N!} \\
&= \frac{n!}{g!b!}\,\frac{G(G-1)\cdots(G-g+1)\;B(B-1)\cdots(B-b+1)}{N(N-1)\cdots(N-b-g+1)} \\
&\approx \binom{n}{g}\left(\frac{G}{N}\right)^{g}\left(\frac{B}{N}\right)^{b}
\end{aligned}
$$

- $\Rightarrow$ sampling with and without replacement yield about the same probabilities for large populations.

Sampling many types



- N items, N_j of type j, $N = \sum_{j=1}^{K} N_j$.
- n items are chosen <u>without</u> replacement.

$$P\{n_j \text{ of type } j,\ j = 1, \ldots, K\ \} \ = \ \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_K}{n_K}}{\binom{N}{n}}$$

for $n = n_1 + n_2 + \ldots + n_K$

- n items are chosen <u>with</u> replacement.

$$\text{Prob} \ = \ \frac{n!}{n_1! n_2! \ldots n_K!}\left(\frac{N_1}{N}\right)^{n_1}\left(\frac{N_2}{N}\right)^{n_2}\cdots\left(\frac{N_K}{N}\right)^{n_K}$$

Other sampling problems

- N items, n are chosen at random with replacement.

$$P\{\text{No repetition in sample}\} = \frac{N(N-1)\cdots(N-n+1)}{N^n} = \frac{(N)_n}{N^n}.$$

e.g. Choose 5 digits from $0, 1, \ldots, 9$ at random. Prob that all are different is $(10)_5 10^{-5} = 0.3024$.

- N items are placed randomly into N cells.

$$P\{\text{Each cell occupied}\} = \frac{N!}{N^N}.$$

Since $7!/7^7 = 0.00612$ in a city where 7 accidents occur each week, practically all weeks will contain days with two or more accidents.

- N balls are placed randomly into $n \geq N$ cells.

$$P\{\text{No cell contains more than one ball}\} = \frac{(n)_N}{n^N}.$$

If n elevator starts with 7 passengers and stops at 10 floors,

$$P\{\text{Everyone leaves at a different floor}\} = \frac{(10)_7}{7^10} = 0.06048.$$

Random variables

- 1: The number of heads in n tosses of a coin.
- 2: The number of cars seen by a wombat.
- 3: The number of wombats killed in a day.
- 4: Time until wombat death.
- 5: The number of "good" (or "bad") items.
- 6: The numbers of items of various kinds.

	1	2	3	4	5	6
Mode	Y	Y	y		y	
Mean (Expectation)	y	y	y	y	y	?
Dependence					Y	y

©Tom Dempsey / Photoseek.com