Lectures prepared by:
Elchanan Mossel
Yelena Shvets

Follows Jim Pitman’s book:
Probability
Section 2.5
Sampling with replacement

Suppose we have a population of size $N$ with $G$ good elements and $B$ bad elements. We draw $n$ times with replacement from this population.

The number $g$ of good elements in the sample will have a binomial$(n,p)$ distribution with $p = G/N$ and $1-p = B/N$.

$$P(g\text{ good and } b\text{ bad}) = \binom{n}{g} \frac{(G)^g (B)^b}{(N)^n}.$$
• If $n$ is large, this will be well approximated by $N(np, \sqrt{np(1-p)})$.

• The proportion of good elements in the sample $g/n$ will lie in the interval $p \pm \frac{1}{\sqrt{n}}$ with probability 95%.

• If the $p$ is not known, it can be estimated by the method of confidence intervals.
Confidence intervals

Suppose we observe the results of $n$ trials with an unknown probability of success $p$.

The observed frequency of successes $\hat{p} = \frac{\text{#successes}}{n}$. 
The Normal Curve Approximation says that for any fixed $p$ and $n$ large enough, there is a 99.99% chance that the observed frequency $\hat{p}$ will differ from $p$ by less than $4\sqrt{p(1-p)/n}$. It's easy to see that $\sqrt{p(1-p)} \leq \frac{1}{2}$, so $4\sqrt{p(1-p)/n} \leq \frac{2}{\sqrt{n}}$. 
is called a 99.99% confidence interval.
Sampling without replacement

• Let’s now think about drawing *without* replacement. The sample size has to be restricted to \( n \leq N \).

• Then number of possible orderings of \( n \) elements out of \( N \) is:

\[
(N)_n = N(N-1)(N-2) \ldots (N-n+1).
\]

• \((N)_n\) is called *N order n*
Sampling without replacement

Note that:

\[
\binom{N}{n} = \frac{(N)_n}{n!}.
\]

So:

\[
(N)_n = \binom{N}{n} n!.
\]
Sampling without replacement

• The chance of getting a specific sample with \( g \) good elements followed by \( b \) bad ones is:

\[
\frac{G}{N} \cdot \frac{G-1}{N-1} \cdot \frac{G-g+1}{N-g+1} \cdot \frac{B}{N-g} \cdot \frac{B-1}{N-g-1} \cdot \frac{B-b+1}{N-g-b+1} = \binom{G}{g} \binom{B}{b} / \binom{N}{n}.
\]

• Since there are \( \binom{n}{g} \) samples with \( g \) good and \( b \) bad elements all having the same probability, we obtain:
Sampling with and without replacement

• For sampling without replacement:

\[
P(\text{g good and b bad}) = \binom{n}{g} \frac{(G)^g (B)^b}{(N)^n} = \frac{\binom{G}{g} \binom{B}{b}}{\binom{N}{n}}.
\]

• For sampling with replacement:

\[
P(\text{g good and b bad}) = \binom{n}{g} \frac{(G)^g (B)^b}{(N)^n}.
\]
Hypergeometric Distribution.

• The distribution of the number of good elements in a sample of
  - size n
  - without replacement
• From a population of
  - G good and
  - N-G = B bad elements
Is called the hypergeometric distribution with parameters \((n,N,G)\).
Sampling with and without replacement

- When $N$ is large, $(N)_n / N^n \to 1$.
- When $B$ is large, $(B)_b / B^b \to 1$.
- When $G$ is large, $(G)_g / G^g \to 1$.

So for fixed $b,g$ and $n$ as $B,G,N \to \infty$ the hypergeometric distribution can be approximated by a binomial($n, G/N$).
Multinomial Distribution

Suppose each trial can result in $m$ possible categories $c_1, c_2, \ldots, c_m$ with probabilities $p_1, p_2, \ldots, p_m$, where $p_1+p_2+\ldots+p_m = 1$.

Suppose we make a sequence of $n$ independent trials and let $N_i$ denote the number of results in the $i^{th}$ category $c_i$. 
Multinomial Distribution

Then for every \( m \)-tuple of non-negative integers \( (n_1, n_2, \ldots, n_m) \) with \( n_1+n_2+\ldots+n_m = n \)

\[
P(N_1=n_1,N_2=n_2,\ldots,N_m=n_m) = \frac{n!}{n_1!n_2!\ldots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m}
\]

Probability of any specific sequence

Number of possible sequences with the given elements
1, 3, 5 and 'even'

Suppose we roll a fair die 10 times and record the number of

\[ \text{ Roll: } 1, 2, 3, \text{ and even's.} \]

**Question:**

What's the probability of seeing

\[ 1, 2, 3, \text{ and 4 even numbers?} \]
Using the multinomial distribution:

\[
P(N_1=1,N_3=2,N_5=3,N_{\text{even}}=4) = \frac{10!}{1!2!3!4!} \left( \frac{1}{6} \right)^1 \left( \frac{1}{6} \right)^2 \left( \frac{1}{6} \right)^3 \left( \frac{3}{6} \right)^4
\]

\[
= 0.016878858
\]
Hypergeometric Extension

• Consider also: this is the no. of ways to choose \( g \) things from \( G \) and \( b \) bad things from \( B \), out of all possible ways to choose \( n \) things from \( N \).

• This way of counting lets us generalize to multiple subcategories easily. How many ways are there to choose \( g \) good from \( G \) and \( b \) bad from \( B \) and \( o \) ok’s from \( O \) and \( p \) passable from \( P \) out of all possible ways to choose \( n \) things from \( N \)?