Hypergeometric

- $\boxed{\text{Urn model:}}$
  - m white balls
  - n black balls
  - k draws without replacement
  - $X \sim \text{Hyper}(m, n, k)$ counts number of white balls drawn
- Derive pmf

$$P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

- Derive $E(X)$, $\text{Var}(X)$

Can think of $X_i = 0$ or 1 based on whether the ith draw is black or white

$$X_i \sim \text{Binom}\left(1, \frac{m}{m+n}\right) \qquad \left(\text{refer } \pi = \frac{m}{m+n}\right)$$

$$E(X) = E(X_1 + \dots + X_k) = k\pi \qquad \left(\text{same as } X \sim \text{Binom}(k, \pi)\right)$$

$$\text{Var}(X) = \underbrace{k\pi(1-\pi)}_{\text{var of binom.}} \cdot \underbrace{\left(\frac{m+n-k}{m+n-1}\right)}_{\text{fudge factor}}$$

fudge factor appears since not replacement

$X \sim \text{Hyper}(15, 10, 5)$

$P(X = 4) =$ prob. of 4 white, 1 black $= \text{dhyper}(4, 15, 10, 5)$

Benford's Law

See p. 103

Let X be the leading digit of some recorded number on a balance sheet, tax return, etc. Consider the pmf(?):

```
log(2:10, 10) - log(1:9, 10)

[1] 0.30103000 0.17609126 0.12493874 0.09691001 0.07918125 0.06694679 0.05799195
[8] 0.05115252 0.04575749
```

Multinomial

The setting is the same as binomial except for these alterations:

- We assume each of the n trials has $k \geq 2$ possible outcomes. In binomial, $k = 2$.
- In binomial settings, π is the probability of "success" and, necessarily, the probability of "failure" is $1 - \pi$. Now we have individual probabilities for each of the k outcomes: π_1 for outcome 1, π_2 for outcome 2, ..., π_k for outcome k. Naturally,

  $$\pi_1 + \pi_2 + \cdots + \pi_k = 1.$$

  When convenient, we will denote this list of probabilities by a vector $\boldsymbol{\pi} = \langle \pi_1, \pi_2, \ldots, \pi_k \rangle$.
- In binomial settings, we counted successes, often denoting this count as X. If $X \sim \text{Binom}(n, \pi)$, then $n - X$ is the number of failures.

  Now, we count occurrences of each of the outcomes: X_1 is the number of times in n trials that outcome 1 occurs, X_2 is the number of times in n trials that outcome 2 occurs, ..., X_k is the number of times in n trials that outcome k occurs. We have

  $$X_1 + X_2 + \cdots + X_k = n,$$

  and will sometimes refer to the full list in vector form $\mathbf{X} = \langle X_1, X_2, \ldots, X_k \rangle$.

The pmf for such a **random vector X** can be derived, yielding

$$P(\mathbf{X} = \mathbf{x}) = \binom{n}{\mathbf{x}} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_k^{x_k} = \binom{n}{x_1 x_2 \cdots x_k} \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_k^{x_k}.$$

This involves new notation for a **multinomial coefficient**

$$\binom{n}{\mathbf{x}} = \binom{n}{x_1 x_2 \cdots x_k} := \binom{n}{x_1} \binom{n - x_1}{x_2} \binom{n - x_1 - x_2}{x_3} \cdots \binom{x_k}{x_k} = \frac{n!}{x_1! x_2! \cdots x_k!}.$$

$H_0: \pi = 0.5$ $\qquad$ $H_a: \underline{\pi > 0.5}$

proportion of times Gus uses his right paw

data: $X = 8$ of 10 trials, were successes

Null $X \sim \text{Binom}(10, 0.5)$

P-value: $P(X=8) + P(X=9) + P(X=10) = 0.055$.

If the alt. hyp. had been 2-sided:

$X=2$ is just as extremely different from $E(X) = 5$ as $X=8$

One approach to P-value

$P(X=0) + P(X=1) + P(X=2) + 0.055 = \text{ans} = 0.11$.

Another approach:

Add up all probabilities that don't exceed $P(X=8)$.

\# 2.57 (a)

$H_0: \pi = 0.5$ $\qquad$ $H_a: \pi \neq 0.5$

$\alpha = 0.05$

X = count of heads in null world $\sim \text{Binom}(200, 0.5)$

Will reject if our count lands in rejection region (where $P < 0.05$)

sum(dbinom(rejectionRegion, 200, 0.55)) = 0.262.