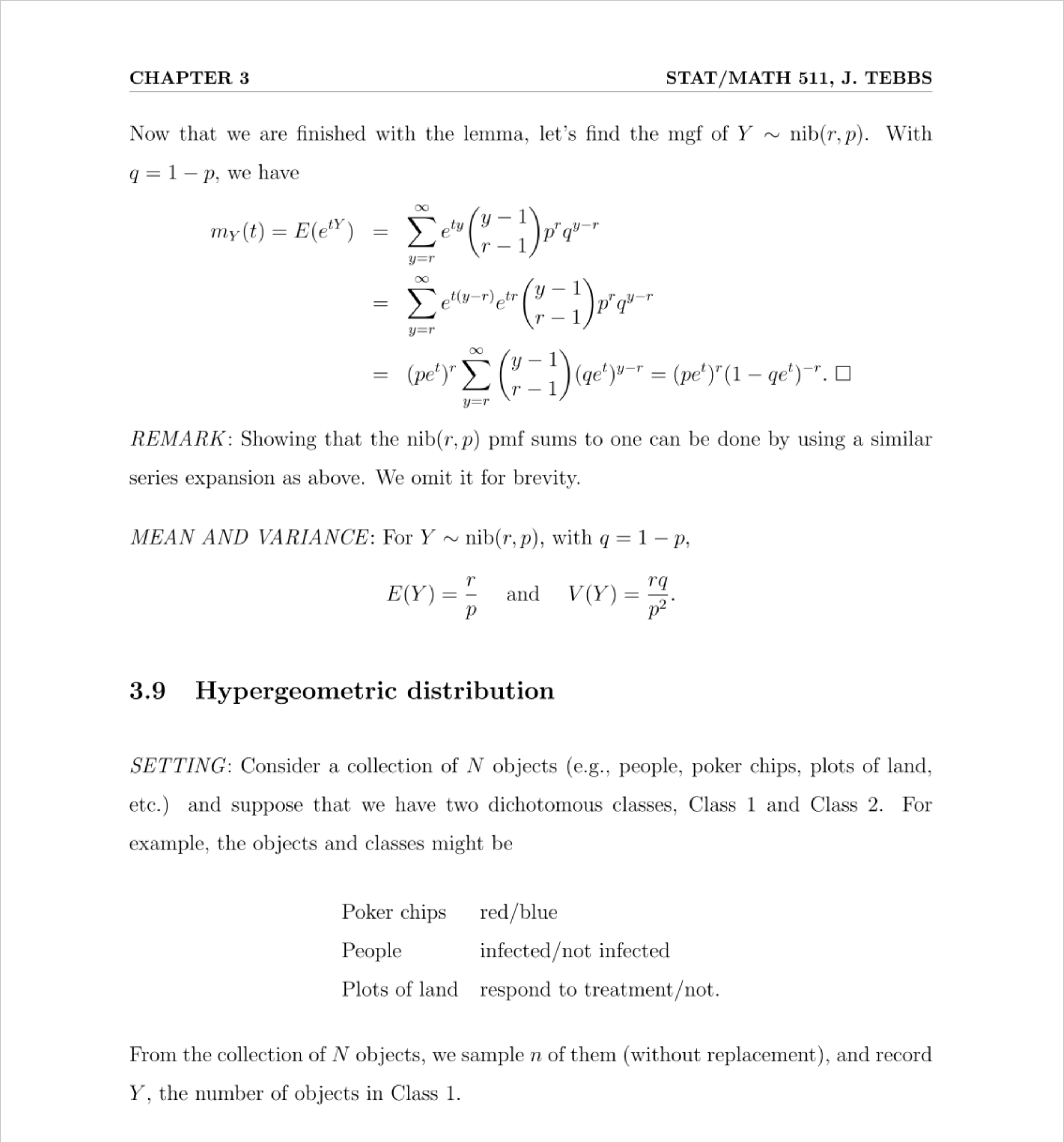
Section 3.9 Hypergeometric Distribution

Tuesday, October 4, 2016 4:06 PM

Section 3.9 Hypergeo...



Now that we are finished with the lemma, let's find the mgf of $Y \sim \text{nib}(r,p)$. With $q = 1 - p$, we have

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y=r}^{\infty} e^{ty} \binom{y-1}{r-1} p^r q^{y-r} \\ &= \sum_{y=r}^{\infty} e^{t(y-r)} e^{tr} \binom{y-1}{r-1} p^r q^{y-r} \\ &= (pe^t)^r \sum_{y=r}^{\infty} \binom{y-1}{r-1} (qe^t)^{y-r} = (pe^t)^r (1 - qe^t)^{-r}. \ \square \end{aligned}$$

REMARK: Showing that the nib(r,p) pmf sums to one can be done by using a similar series expansion as above. We omit it for brevity.

MEAN AND VARIANCE: For $Y \sim \text{nib}(r,p)$, with $q = 1 - p$,

$$E(Y) = \frac{r}{p} \quad \text{and} \quad V(Y) = \frac{rq}{p^2}.$$

3.9 Hypergeometric distribution

SETTING: Consider a collection of N objects (e.g., people, poker chips, plots of land, etc.) and suppose that we have two dichotomous classes, Class 1 and Class 2. For example, the objects and classes might be

Poker chips	red/blue
People	infected/not infected
Plots of land	respond to treatment/not.

From the collection of N objects, we sample n of them (without replacement), and record Y, the number of objects in Class 1.

Y, the number of objects in Class 1.

REMARK: This sounds like a binomial setup! However, the difference here is that N, the **population size**, is finite (the population size, theoretically, is assumed to be infinite in the binomial model). Thus, if we sample from a population of objects **without replacement**, the "success" probability changes from trial to trial. This, violates the binomial

model assumptions! If N is large (i.e., in a very large population), the hypergeometric and binomial models will be similar, because the change in the probability of success from trial to trial will be small (maybe so small that it is not of practical concern).

HYPERGEOMETRIC DISTRIBUTION: Envision a collection of n objects sampled (at random and without replacement) from a population of size N, where r denotes the size of Class 1 and $N - r$ denotes the size of Class 2. Let Y denote the number of objects in the sample that belong to Class 1. Then, Y has a **hypergeometric distribution**, written $Y \sim \text{hyper}(N, n, r)$, where

$$\begin{aligned} N &= \text{total number of objects} \\ r &= \text{number of the 1st class (e.g., ``success'')} \\ N - r &= \text{number of the 2nd class (e.g., ``failure'')} \\ n &= \text{number of objects sampled.} \end{aligned}$$

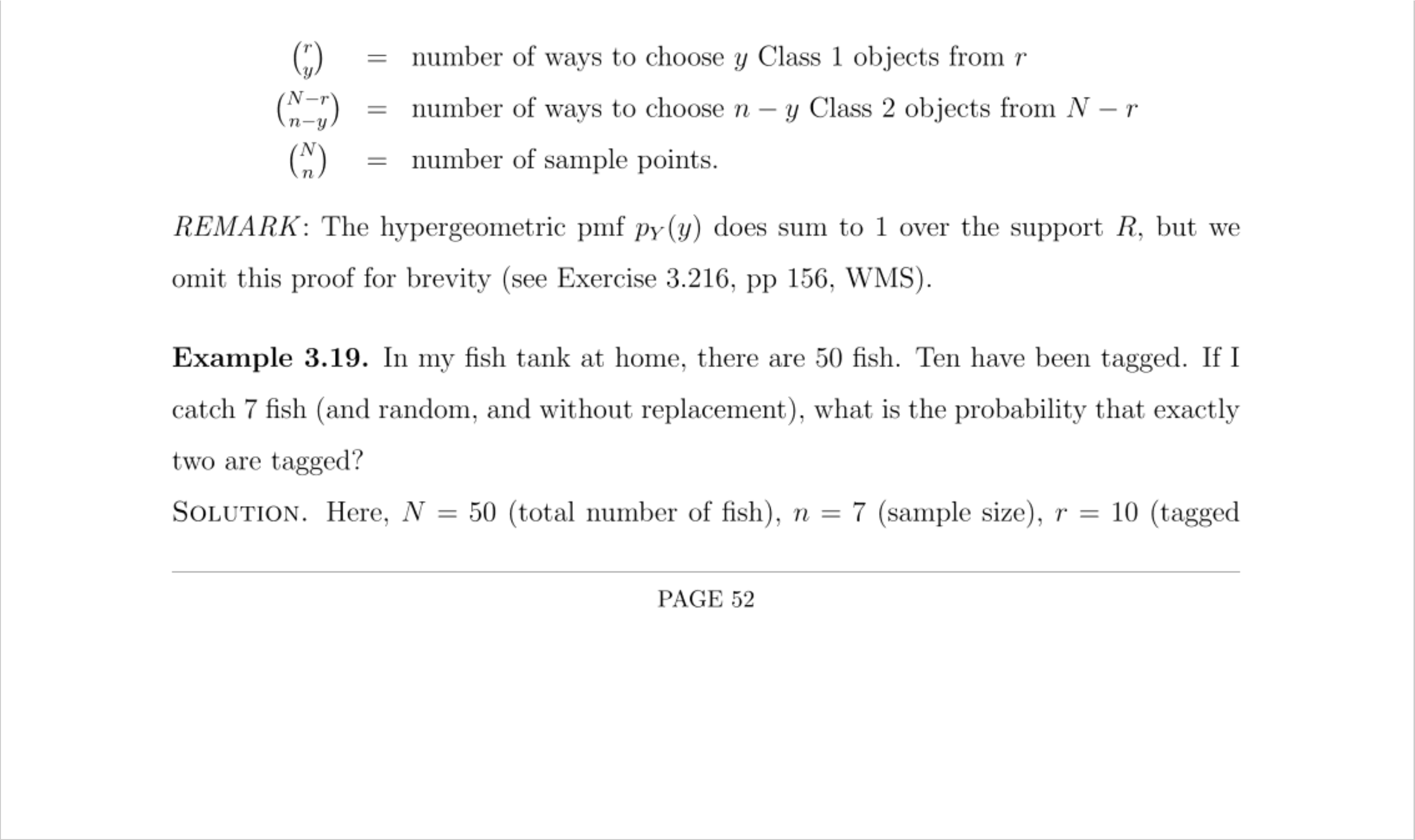
HYPERGEOMETRIC PMF: The pmf for $Y \sim \text{hyper}(N, n, r)$ is given by

$$p_Y(y) = \begin{cases} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, & y \in R \\ 0, & \text{otherwise}, \end{cases}$$

where the support set $R = \{y \in \mathcal{N} : \max(0, n - N + r) \le y \le \min(n, r)\}$.

BREAKDOWN: In the hyper(N, n, r) pmf, we have three parts:

$$\binom{r}{y} = \text{number of ways to choose } y \text{ Class 1 objects from } r$$

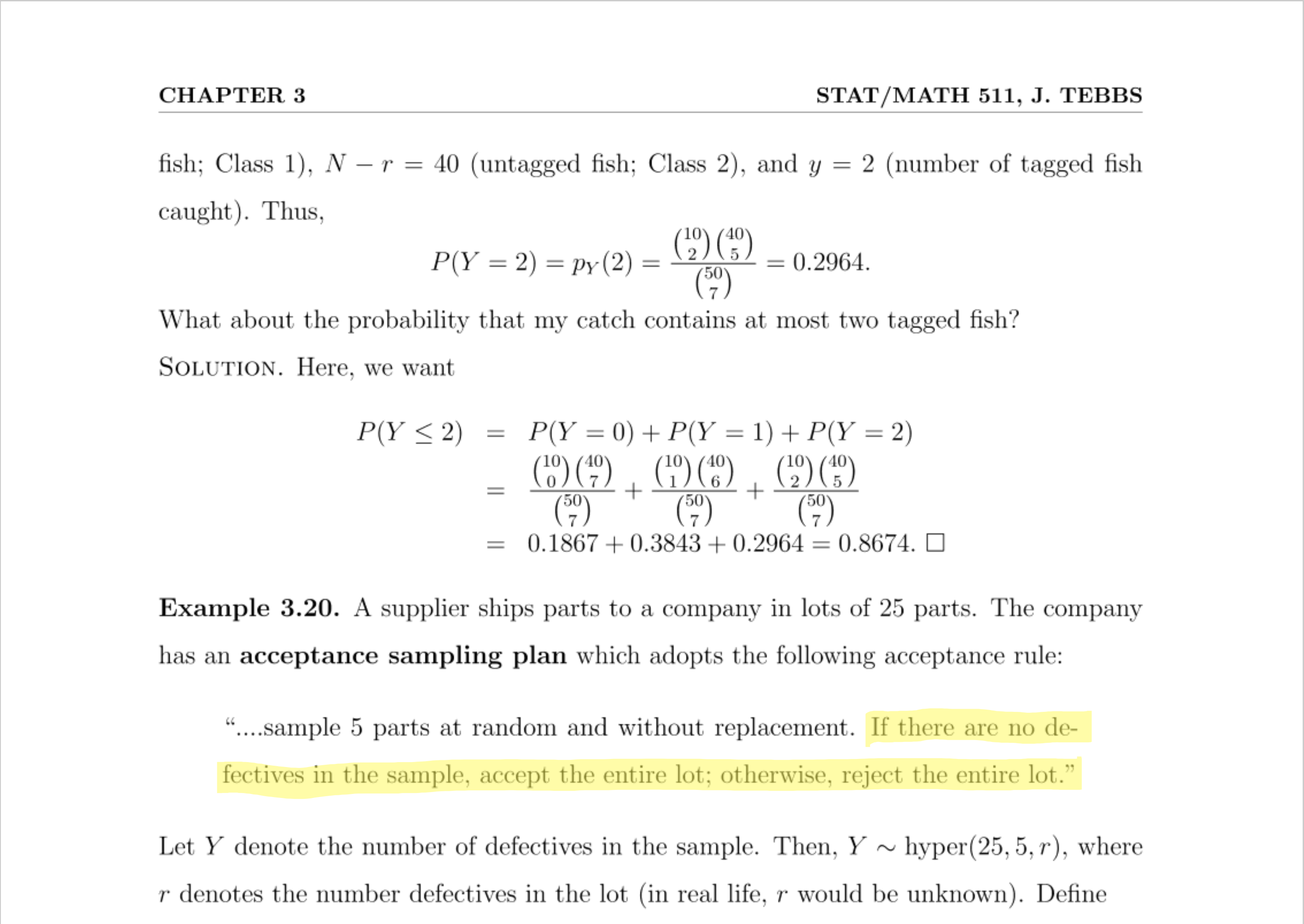


$$\begin{aligned}\binom{r}{y} &= \text{number of ways to choose } y \text{ Class 1 objects from } r\\ \binom{N-r}{n-y} &= \text{number of ways to choose } n-y \text{ Class 2 objects from } N-r\\ \binom{N}{n} &= \text{number of sample points.}\end{aligned}$$

REMARK: The hypergeometric pmf $p_Y(y)$ does sum to 1 over the support R, but we omit this proof for brevity (see Exercise 3.216, pp 156, WMS).

Example 3.19. In my fish tank at home, there are 50 fish. Ten have been tagged. If I catch 7 fish (and random, and without replacement), what is the probability that exactly two are tagged?

SOLUTION. Here, $N = 50$ (total number of fish), $n = 7$ (sample size), $r = 10$ (tagged



fish; Class 1), $N - r = 40$ (untagged fish; Class 2), and $y = 2$ (number of tagged fish caught). Thus,

$$P(Y = 2) = p_Y(2) = \frac{\binom{10}{2}\binom{40}{5}}{\binom{50}{7}} = 0.2964.$$

What about the probability that my catch contains at most two tagged fish?

SOLUTION. Here, we want

$$\begin{aligned}P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2)\\ &= \frac{\binom{10}{0}\binom{40}{7}}{\binom{50}{7}} + \frac{\binom{10}{1}\binom{40}{6}}{\binom{50}{7}} + \frac{\binom{10}{2}\binom{40}{5}}{\binom{50}{7}}\\ &= 0.1867 + 0.3843 + 0.2964 = 0.8674. \ \square\end{aligned}$$

Example 3.20. A supplier ships parts to a company in lots of 25 parts. The company has an **acceptance sampling plan** which adopts the following acceptance rule:

> "....sample 5 parts at random and without replacement. If there are no defectives in the sample, accept the entire lot; otherwise, reject the entire lot."

Let Y denote the number of defectives in the sample. Then, $Y \sim \text{hyper}(25, 5, r)$, where r denotes the number defectives in the lot (in real life, r would be unknown). Define

Let Y denote the number of defectives in the sample. Then, $Y \sim \text{hyper}(25, 5, r)$, where r denotes the number defectives in the lot (in real life, r would be unknown). Define

$$OC(p) = P(Y = 0) = \frac{\binom{r}{0}\binom{25-r}{5}}{\binom{25}{5}},$$

where $p = r/25$ denotes the true proportion of defectives in the lot. The symbol $OC(p)$ denotes the probability of accepting the lot (which is a function of p). Consider the following table, whose entries are computed using the above probability expression:

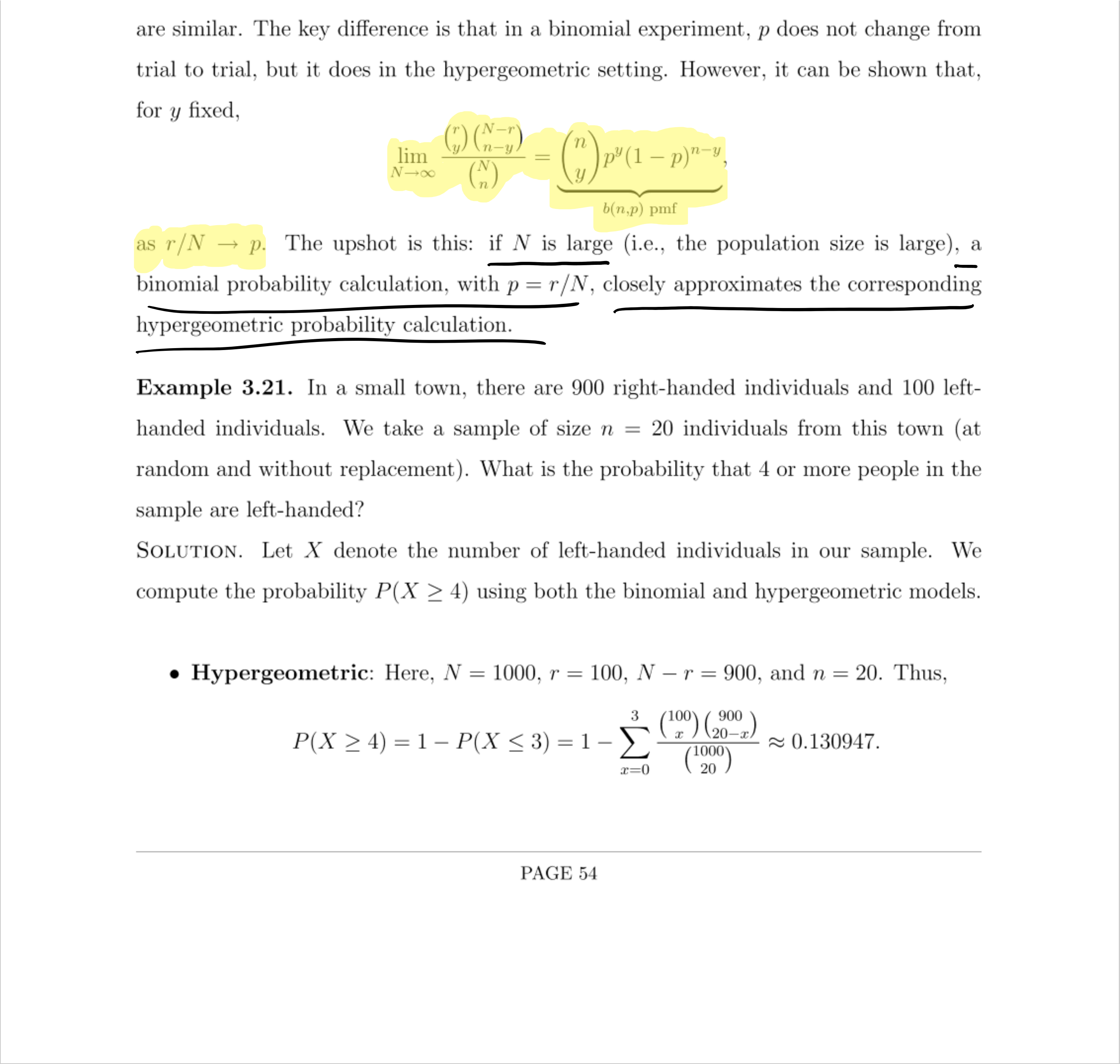
r	p	$OC(p)$
0	0	1.00
1	0.04	0.80
2	0.08	0.63
3	0.12	0.50
4	0.16	0.38
5	0.20	0.29
10	0.40	0.06
15	0.60	0.01

REMARK: The graph of $OC(p)$ versus p is called an **operating characteristic curve**. For sensible sampling plans, $OC(p)$ is a decreasing function of p. Acceptance sampling is an important part of **statistical process control**, which is used in engineering and manufacturing settings. □

MEAN AND VARIANCE: If $Y \sim \text{hyper}(N, n, r)$, then

$$\begin{aligned} E(Y) &= n\left(\frac{r}{N}\right) \\ V(Y) &= n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right). \end{aligned}$$

RELATIONSHIP WITH THE BINOMIAL: The binomial and hypergeometric models are similar. The key difference is that in a binomial experiment, p does not change from trial to trial, but it does in the hypergeometric setting. However, it can be shown that,



are similar. The key difference is that in a binomial experiment, p does not change from trial to trial, but it does in the hypergeometric setting. However, it can be shown that, for y fixed,

$$\lim_{N\to\infty}\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \underbrace{\binom{n}{y}p^y(1-p)^{n-y}}_{b(n,p)\text{ pmf}},$$

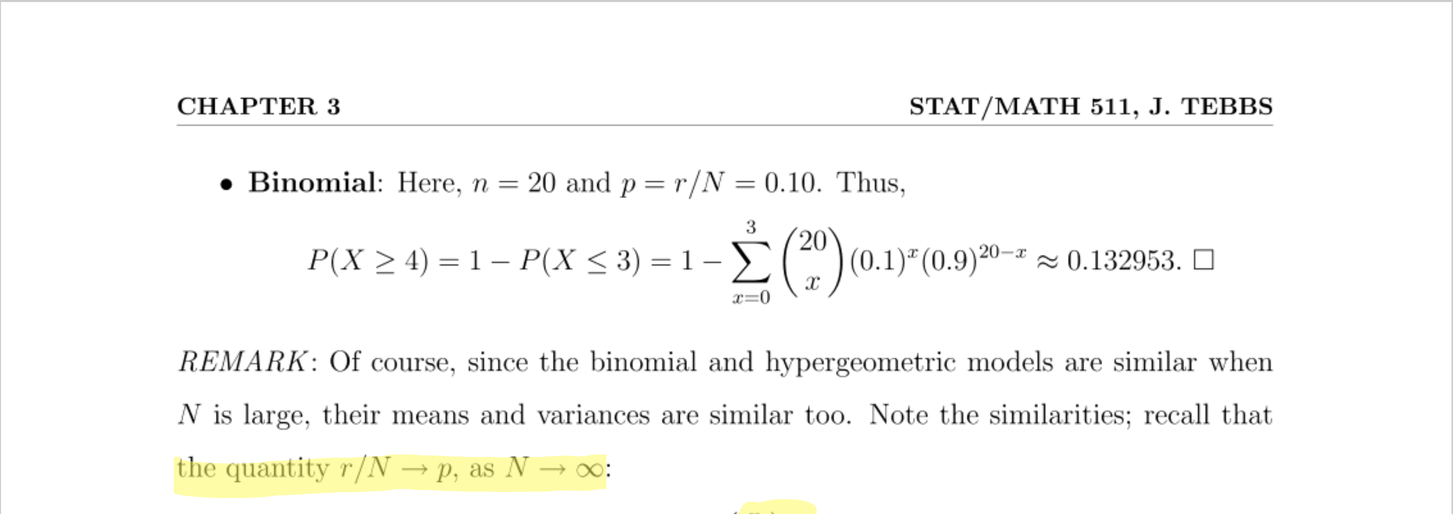
as $r/N \to p$. The upshot is this: if N is large (i.e., the population size is large), a binomial probability calculation, with $p = r/N$, closely approximates the corresponding hypergeometric probability calculation.

Example 3.21. In a small town, there are 900 right-handed individuals and 100 left-handed individuals. We take a sample of size $n = 20$ individuals from this town (at random and without replacement). What is the probability that 4 or more people in the sample are left-handed?

SOLUTION. Let X denote the number of left-handed individuals in our sample. We compute the probability $P(X \geq 4)$ using both the binomial and hypergeometric models.

- **Hypergeometric**: Here, $N = 1000$, $r = 100$, $N - r = 900$, and $n = 20$. Thus,

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3}\frac{\binom{100}{x}\binom{900}{20-x}}{\binom{1000}{20}} \approx 0.130947.$$



- **Binomial**: Here, $n = 20$ and $p = r/N = 0.10$. Thus,

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3}\binom{20}{x}(0.1)^x(0.9)^{20-x} \approx 0.132953. \ \square$$

REMARK: Of course, since the binomial and hypergeometric models are similar when N is large, their means and variances are similar too. Note the similarities; recall that the quantity $r/N \to p$, as $N \to \infty$:

N is large, then means and variances are similar too. Note the similarities, recall that the quantity $r/N \to p$, as $N \to \infty$:

$$E(Y) = n\left(\frac{r}{N}\right) \approx np$$

and

$$V(Y) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right) \approx np(1-p).$$

3.10 Poisson distribution

TERMINOLOGY: Let the number of occurrences in a given continuous interval of time or space be counted. A **Poisson process** enjoys the following properties:

(1) the number of occurrences in non-overlapping intervals are **independent** random variables.

(2) The probability of an occurrence in a sufficiently short interval is **proportional to the length** of the interval.

(3) The probability of 2 or more occurrences in a sufficiently short interval is zero.

GOAL: Suppose that a process satisfies the above three conditions, and let Y denote the number of occurrences in an interval of length one. Our goal is to find an expression for $p_Y(y) = P(Y = y)$, the pmf of Y.

APPROACH: Envision partitioning the unit interval $[0, 1]$ into n subintervals, each of size $1/n$. Now, if n is sufficiently large (i.e., much larger than y), then we can approximate the probability that y events occur in this unit interval by finding the probability that exactly one event (occurrence) occurs in exactly y of the subintervals.