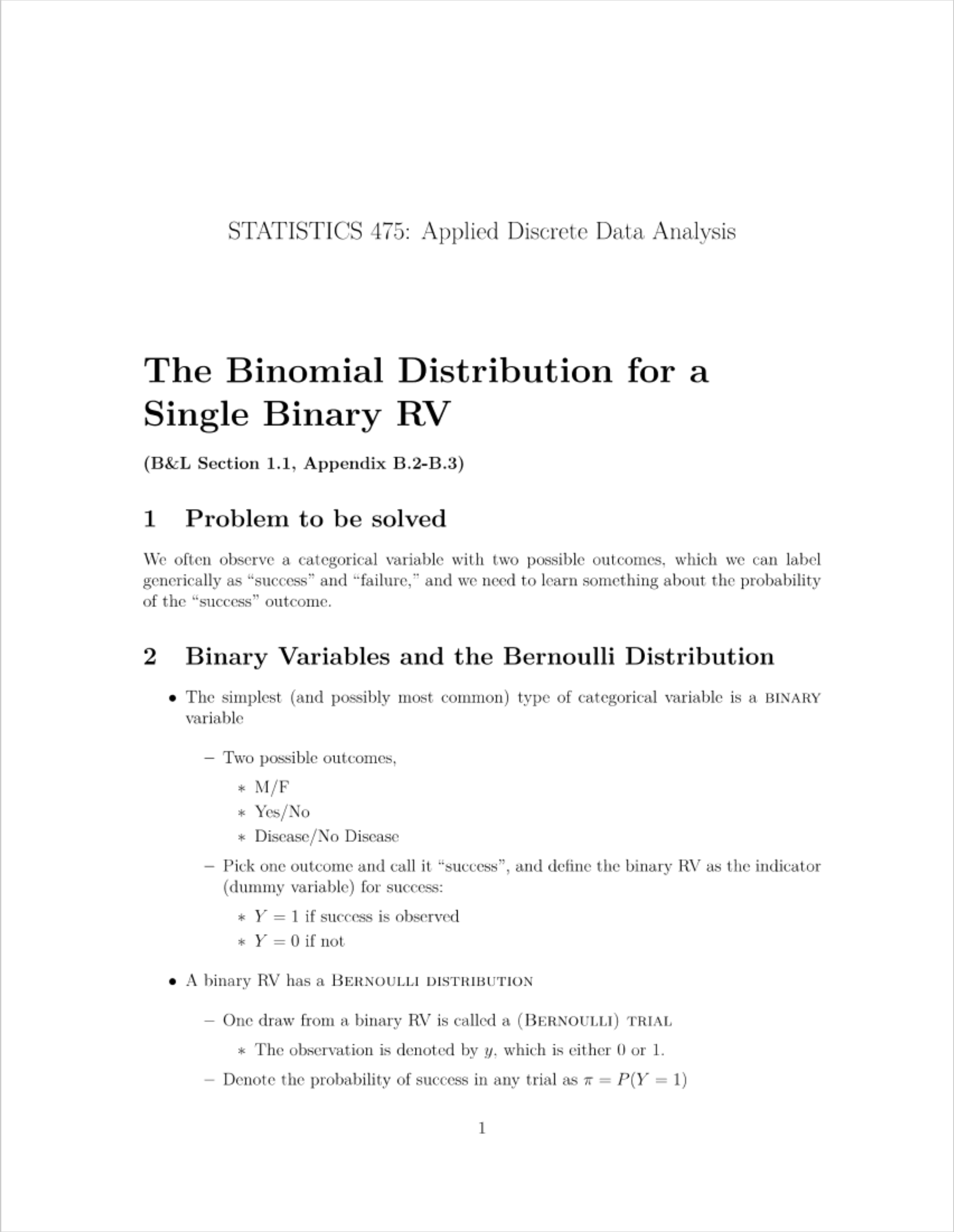
Lecture 2

January 3, 2019 10:54 AM



STATISTICS 475: Applied Discrete Data Analysis

The Binomial Distribution for a Single Binary RV

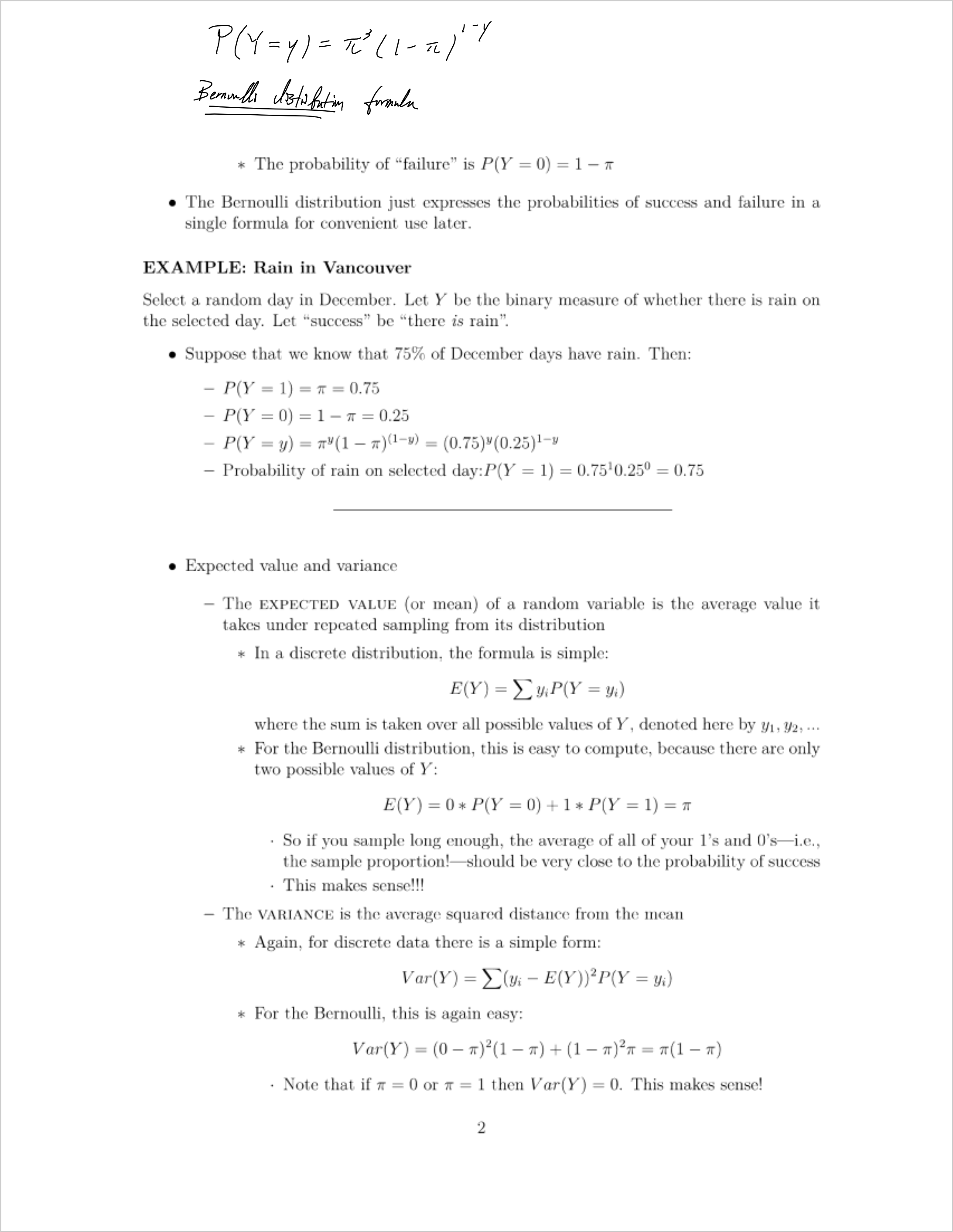
(B&L Section 1.1, Appendix B.2-B.3)

1 Problem to be solved

We often observe a categorical variable with two possible outcomes, which we can label generically as "success" and "failure," and we need to learn something about the probability of the "success" outcome.

2 Binary Variables and the Bernoulli Distribution

- The simplest (and possibly most common) type of categorical variable is a BINARY variable
  - Two possible outcomes,
    - M/F
    - Yes/No
    - Disease/No Disease
  - Pick one outcome and call it "success", and define the binary RV as the indicator (dummy variable) for success:
    - $Y = 1$ if success is observed
    - $Y = 0$ if not
- A binary RV has a BERNOULLI DISTRIBUTION
  - One draw from a binary RV is called a (BERNOULLI) TRIAL
    - The observation is denoted by y, which is either 0 or 1.
  - Denote the probability of success in any trial as $\pi = P(Y = 1)$



$P(Y=y) = \pi^y (1-\pi)^{1-y}$

Bernoulli distribution formula

- * The probability of "failure" is $P(Y=0) = 1 - \pi$

- The Bernoulli distribution just expresses the probabilities of success and failure in a single formula for convenient use later.

EXAMPLE: Rain in Vancouver

Select a random day in December. Let Y be the binary measure of whether there is rain on the selected day. Let "success" be "there *is* rain".

- Suppose that we know that 75% of December days have rain. Then:
  - $P(Y=1) = \pi = 0.75$
  - $P(Y=0) = 1 - \pi = 0.25$
  - $P(Y=y) = \pi^y (1-\pi)^{(1-y)} = (0.75)^y (0.25)^{1-y}$
  - Probability of rain on selected day:$P(Y=1) = 0.75^1 0.25^0 = 0.75$

- Expected value and variance
  - The EXPECTED VALUE (or mean) of a random variable is the average value it takes under repeated sampling from its distribution
    - * In a discrete distribution, the formula is simple:

$$E(Y) = \sum y_i P(Y = y_i)$$

where the sum is taken over all possible values of Y, denoted here by $y_1, y_2, \ldots$

    - * For the Bernoulli distribution, this is easy to compute, because there are only two possible values of Y:

$$E(Y) = 0 * P(Y=0) + 1 * P(Y=1) = \pi$$

      - · So if you sample long enough, the average of all of your 1's and 0's—i.e., the sample proportion!—should be very close to the probability of success
      - · This makes sense!!!
  - The VARIANCE is the average squared distance from the mean
    - * Again, for discrete data there is a simple form:

$$Var(Y) = \sum (y_i - E(Y))^2 P(Y = y_i)$$

    - * For the Bernoulli, this is again easy:

$$Var(Y) = (0-\pi)^2 (1-\pi) + (1-\pi)^2 \pi = \pi(1-\pi)$$

      - · Note that if $\pi = 0$ or $\pi = 1$ then $Var(Y) = 0$. This makes sense!

3 Multiple Bernoulli trials: The Binomial Distribution

- Much of the time we measure multiple Bernoulli trials and summarize the successes and failures into counts
  - n is the total number of trials
  - In trial i, draw from independent and identically distributed Bernoulli trials (same π), denoted by Y_i, $i = 1, \ldots, n$
    - Observe y_i, $i = 1, \ldots, n$
  - Create a new RV, which is the count of successes $W = \sum_{i=1}^{n} Y_i$
    - Observed version is $w = \sum_{i=1}^{n} y_i$

EXAMPLE: Sex of newborns in Canada

Are more girls being born than boys? 1000 babies are randomly sampled from among those born in Canada in the 2000s. Their sex is recorded:

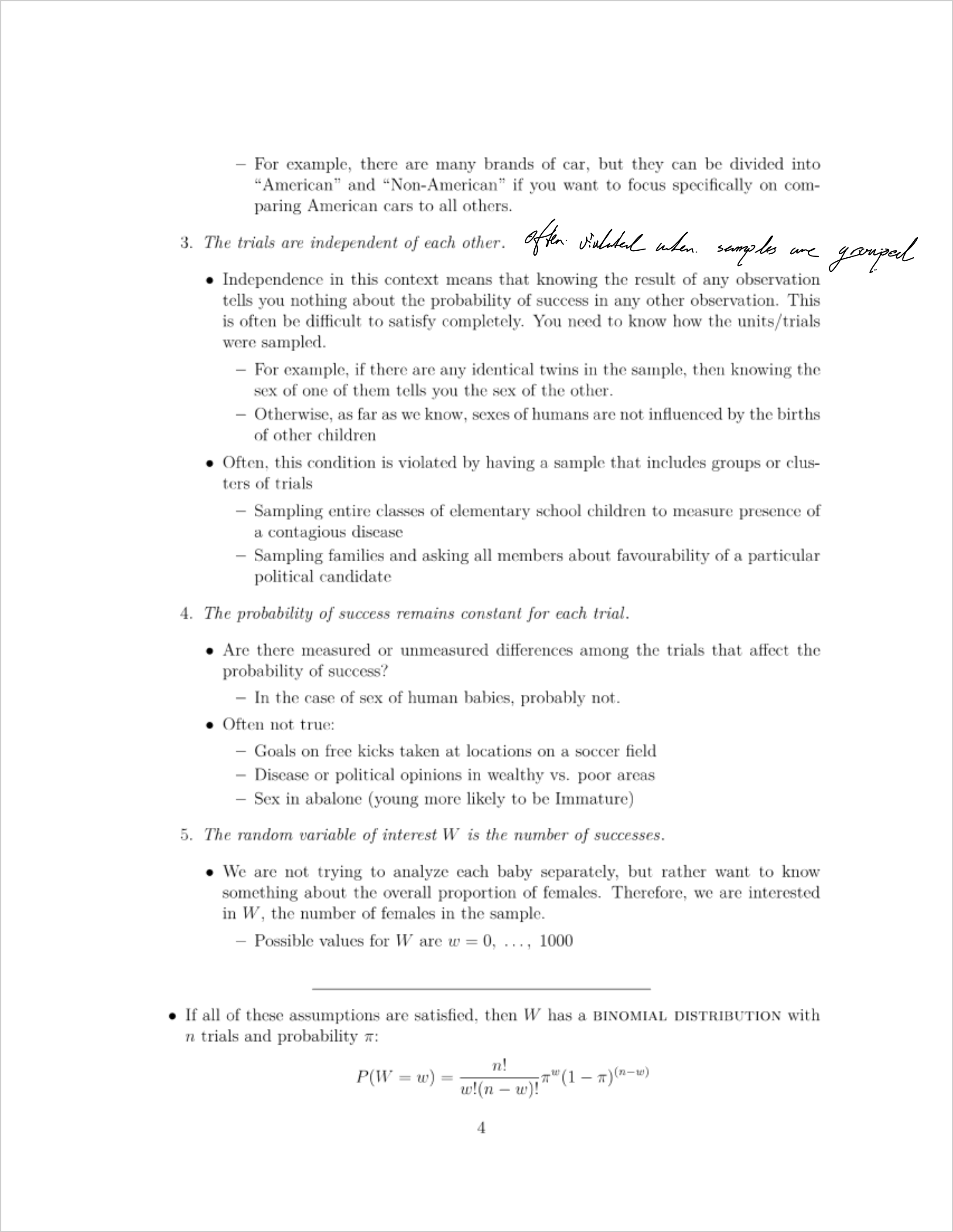
524 Females	476 Males

Define Success=Female, so $\pi = P(\text{Female})$.

- The random variable W—the count of the number of successes in n trials—has a BINOMIAL DISTRIBUTION if the following conditions are met:

Properties of Bernulli

1. *There are n identical trials, or replicates of the same process.*
   - In our example, we have $n = 1000$ babies. They are all human, they are born in Canada...we don't know anything else about them. There are probably different locations, different races of babies, different ages of parents, and so forth. But the *process* that leads to a baby being born and its sex being recorded is essentially the same in every case, so it seems reasonable to assume that they are "identical" trials for the purpose of recording sex.
2. *There are two possible outcomes for each trial.*
   - There are only two possible sexes, and our ability to distinguish them is perfect (sort of...), so this is a reasonable assumption.
   - This is not always the case, however. For example, in some animals (e.g. abalone) sex organs develop later, so a randomly sampled abalone could be F, M, or "Immature".
   - *Many* other categorical variables have more than two possible outcomes. In some cases, however, questions are phrased that suggest two *groups* of possible outcomes.



- For example, there are many brands of car, but they can be divided into "American" and "Non-American" if you want to focus specifically on comparing American cars to all others.

3. *The trials are independent of each other*. Often violated when samples are grouped EX: twins

   - Independence in this context means that knowing the result of any observation tells you nothing about the probability of success in any other observation. This is often be difficult to satisfy completely. You need to know how the units/trials were sampled.
     - For example, if there are any identical twins in the sample, then knowing the sex of one of them tells you the sex of the other.
     - Otherwise, as far as we know, sexes of humans are not influenced by the births of other children
   - Often, this condition is violated by having a sample that includes groups or clusters of trials
     - Sampling entire classes of elementary school children to measure presence of a contagious disease
     - Sampling families and asking all members about favourability of a particular political candidate

4. *The probability of success remains constant for each trial.*

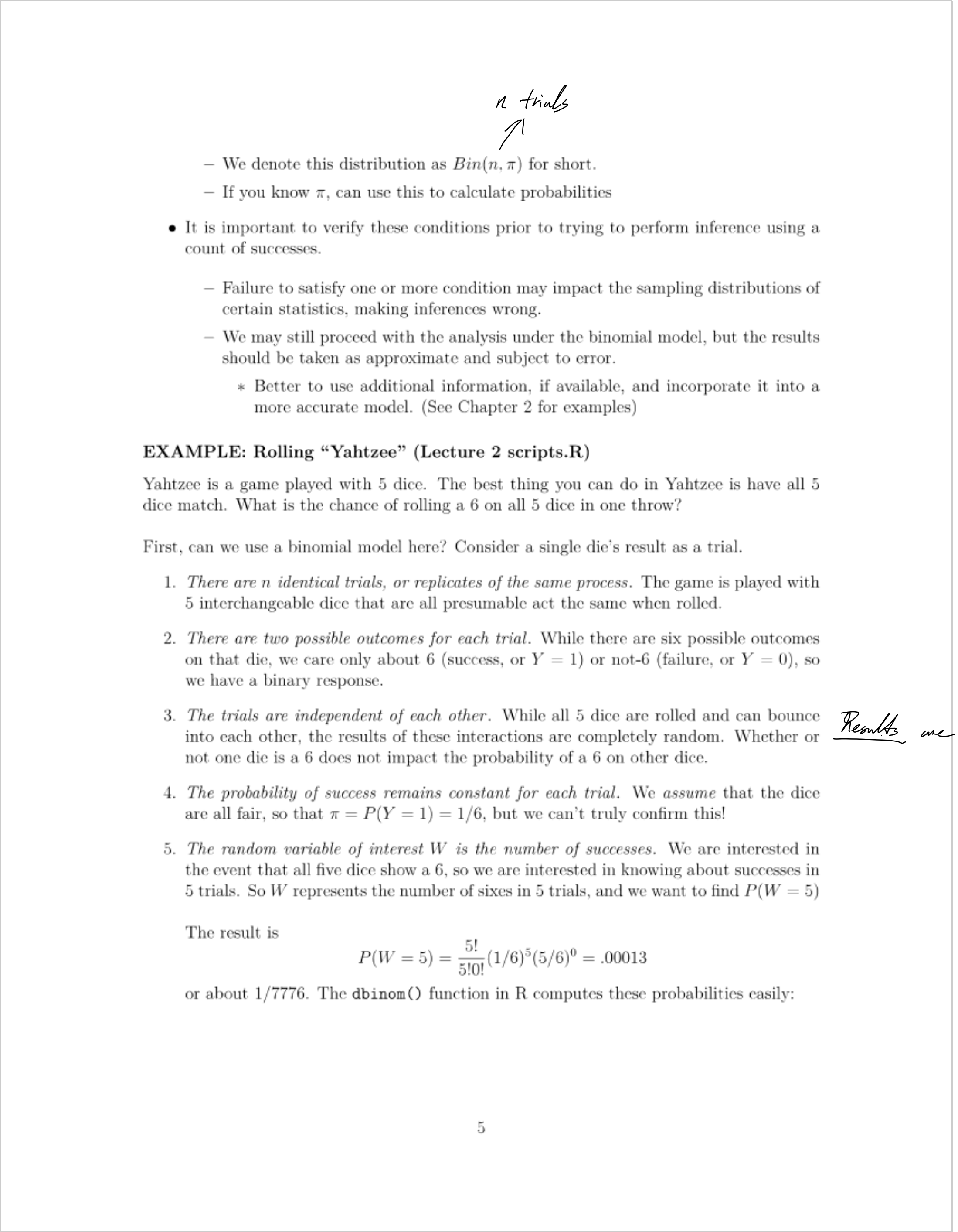
   - Are there measured or unmeasured differences among the trials that affect the probability of success?
     - In the case of sex of human babies, probably not.
   - Often not true:
     - Goals on free kicks taken at locations on a soccer field
     - Disease or political opinions in wealthy vs. poor areas
     - Sex in abalone (young more likely to be Immature)

5. *The random variable of interest W is the number of successes.*

   - We are not trying to analyze each baby separately, but rather want to know something about the overall proportion of females. Therefore, we are interested in W, the number of females in the sample.
     - Possible values for W are $w = 0, \ldots, 1000$

- If all of these assumptions are satisfied, then W has a BINOMIAL DISTRIBUTION with n trials and probability π:

$$P(W = w) = \frac{n!}{w!(n-w)!}\pi^w(1-\pi)^{(n-w)}$$



n trials

- We denote this distribution as $Bin(n, \pi)$ for short.
- If you know π, can use this to calculate probabilities

- It is important to verify these conditions prior to trying to perform inference using a count of successes.
  - Failure to satisfy one or more condition may impact the sampling distributions of certain statistics, making inferences wrong.
  - We may still proceed with the analysis under the binomial model, but the results should be taken as approximate and subject to error.
    - Better to use additional information, if available, and incorporate it into a more accurate model. (See Chapter 2 for examples)

EXAMPLE: Rolling "Yahtzee" (Lecture 2 scripts.R)

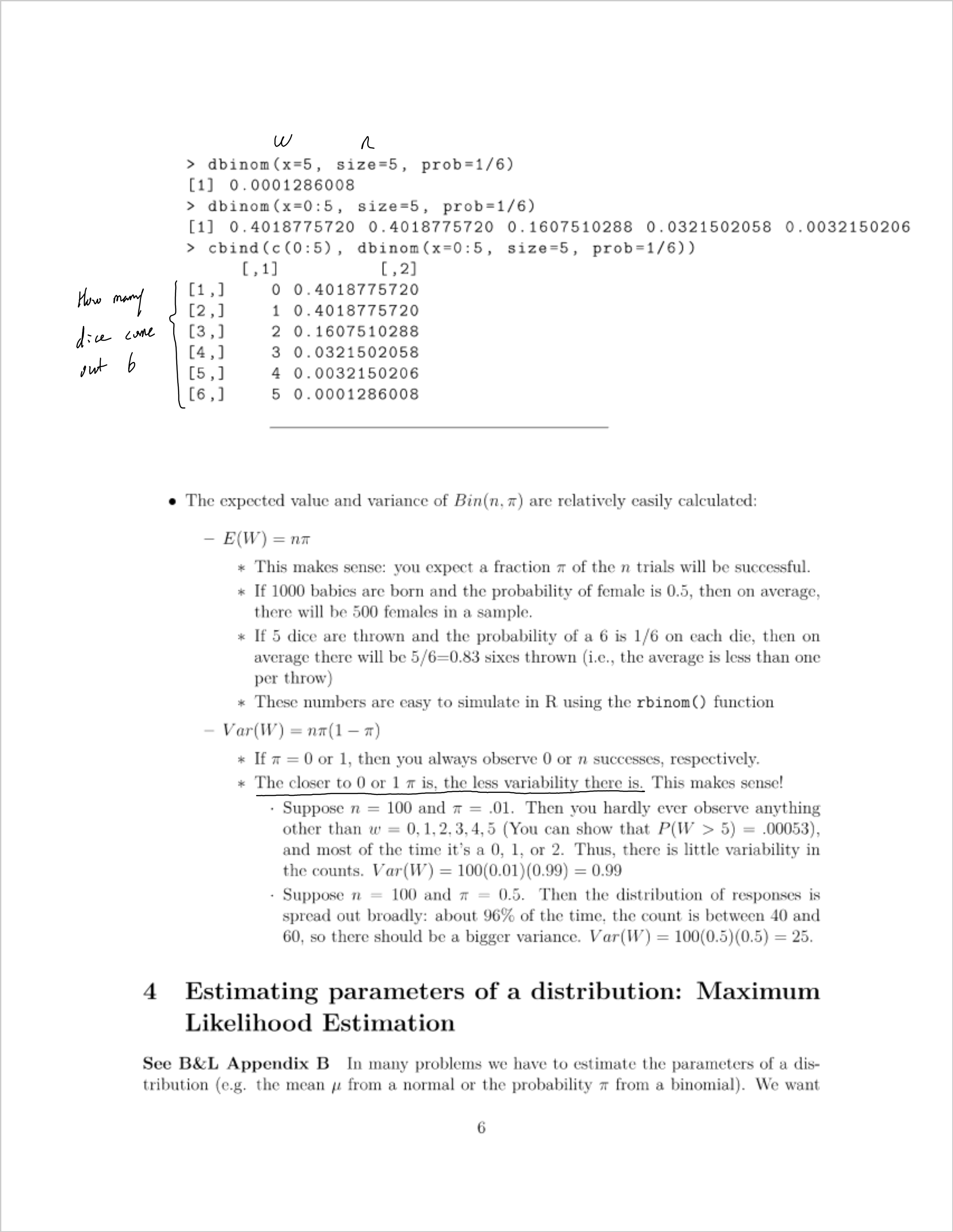
Yahtzee is a game played with 5 dice. The best thing you can do in Yahtzee is have all 5 dice match. What is the chance of rolling a 6 on all 5 dice in one throw?

First, can we use a binomial model here? Consider a single die's result as a trial.

1. *There are n identical trials, or replicates of the same process.* The game is played with 5 interchangeable dice that are all presumable act the same when rolled.
2. *There are two possible outcomes for each trial.* While there are six possible outcomes on that die, we care only about 6 (success, or $Y = 1$) or not-6 (failure, or $Y = 0$), so we have a binary response.
3. *The trials are independent of each other.* While all 5 dice are rolled and can bounce into each other, the results of these interactions are completely random. Whether or not one die is a 6 does not impact the probability of a 6 on other dice.

   Results are independent.
4. *The probability of success remains constant for each trial.* We *assume* that the dice are all fair, so that $\pi = P(Y = 1) = 1/6$, but we can't truly confirm this!
5. *The random variable of interest W is the number of successes.* We are interested in the event that all five dice show a 6, so we are interested in knowing about successes in 5 trials. So W represents the number of sixes in 5 trials, and we want to find $P(W = 5)$

   The result is
   $$P(W = 5) = \frac{5!}{5!0!}(1/6)^5(5/6)^0 = .00013$$
   or about 1/7776. The dbinom() function in R computes these probabilities easily:



w n

```
> dbinom(x=5, size=5, prob=1/6)
[1] 0.0001286008
> dbinom(x=0:5, size=5, prob=1/6)
[1] 0.4018775720 0.4018775720 0.1607510288 0.0321502058 0.0032150206
> cbind(c(0:5), dbinom(x=0:5, size=5, prob=1/6))
     [,1]         [,2]
[1,]    0 0.4018775720
[2,]    1 0.4018775720
[3,]    2 0.1607510288
[4,]    3 0.0321502058
[5,]    4 0.0032150206
[6,]    5 0.0001286008
```

How many dice come out 6

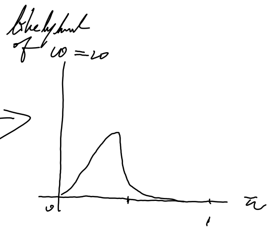
- The expected value and variance of $Bin(n, \pi)$ are relatively easily calculated:
  - $E(W) = n\pi$
    - This makes sense: you expect a fraction π of the n trials will be successful.
    - If 1000 babies are born and the probability of female is 0.5, then on average, there will be 500 females in a sample.
    - If 5 dice are thrown and the probability of a 6 is 1/6 on each die, then on average there will be 5/6=0.83 sixes thrown (i.e., the average is less than one per throw)
    - These numbers are easy to simulate in R using the `rbinom()` function
  - $Var(W) = n\pi(1 - \pi)$
    - If $\pi = 0$ or 1, then you always observe 0 or n successes, respectively.
    - The closer to 0 or 1 π is, the less variability there is. This makes sense!
      - Suppose $n = 100$ and $\pi = .01$. Then you hardly ever observe anything other than $w = 0, 1, 2, 3, 4, 5$ (You can show that $P(W > 5) = .00053$), and most of the time it's a 0, 1, or 2. Thus, there is little variability in the counts. $Var(W) = 100(0.01)(0.99) = 0.99$
      - Suppose $n = 100$ and $\pi = 0.5$. Then the distribution of responses is spread out broadly: about 96% of the time, the count is between 40 and 60, so there should be a bigger variance. $Var(W) = 100(0.5)(0.5) = 25$.

4 Estimating parameters of a distribution: Maximum Likelihood Estimation

See B&L Appendix B In many problems we have to estimate the parameters of a distribution (e.g. the mean μ from a normal or the probability π from a binomial). We want the estimate we compute to be a good one; in fact, we often estimate parameters by defining a reasonable criterion and compute the "best" estimate according to that criterion. (We did this in regression with least squares estimates that minimize the sum of squared errors). Def → One such method of finding "best" estimates according to some criterion. and which we will use extensively for estimating the parameters associated with the models for categorical responses, is called MAXIMUM LIKELIHOOD (ML) ESTIMATION, and the resulting estimate is called the MAXIMUM LIKELIHOOD ESTIMATE (MLE).

If you are not familiar with ML estimation, *read Appendix B*, mainly for the ideas presented in sections B.1, B.2, and B.3. I will show a quick overview here, but there is more to it that what I will do.

EXAMPLE: Binomial probability parameter π Find π the maximizes $P(W = \overset{20}{w})$ →



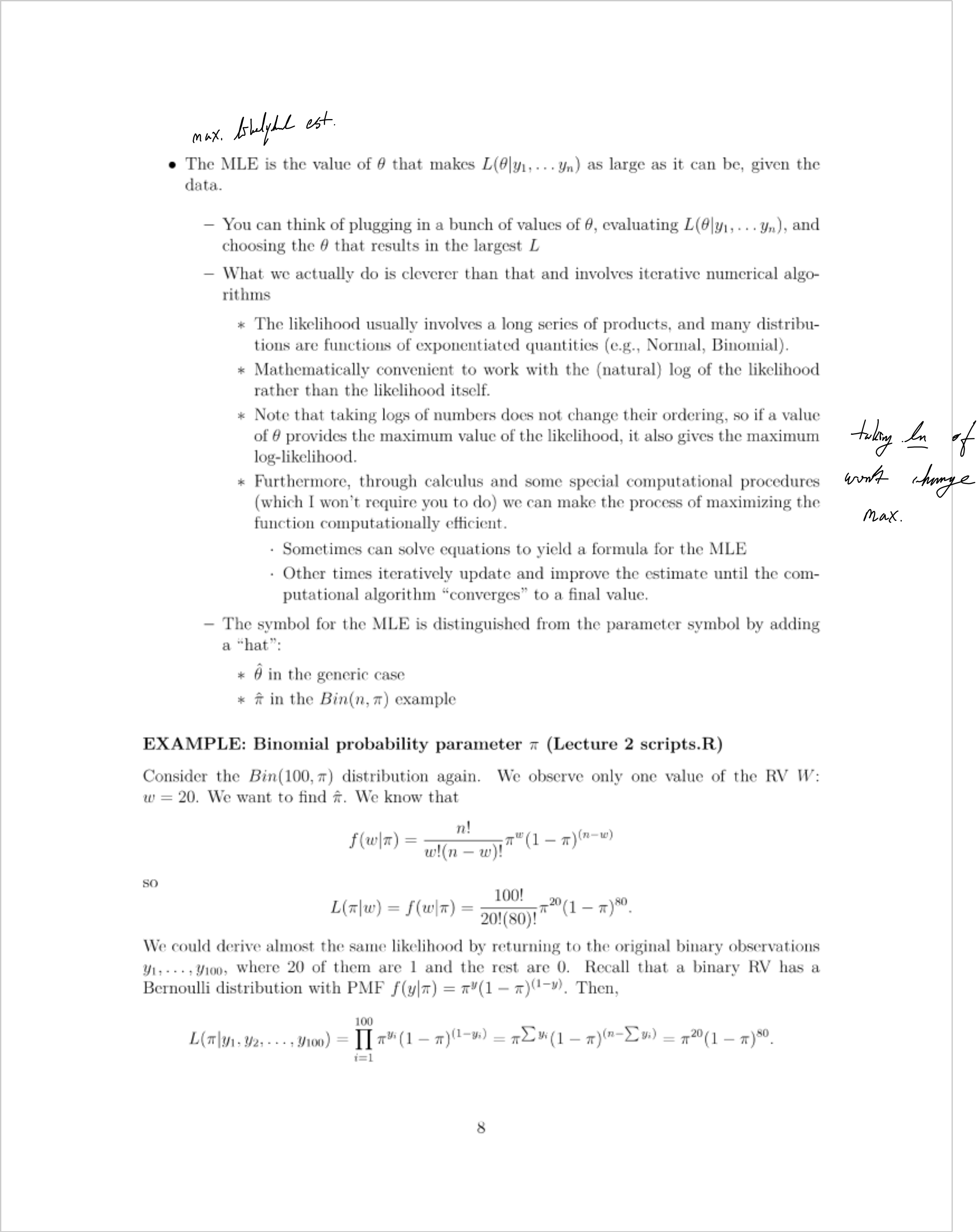
Consider the $Bin(100, \pi)$ distribution. Suppose we observe $w = 20$, but we don't know what the π was that led to that w. We can easily rule out some values because the observed w is unlikely to occur with them. For example, we know that $\pi = 0$ and $\pi = 1$ are impossible. But what about other values?

We could try plugging in a lot of different values of π and picking the one for which $P(W = 20)$ turns out to be the highest. This is exactly what I've done in Figure 1! This is the basis of ML estimation.

- A LIKELIHOOD is a mathematical function describing something similar to the probability of an observed *sample* (as opposed to the probability of a single particular *observation*)
  fix the value of parameter.
  - Generically, let $\underset{w}{Y}$ be a random variable from which we draw our samples
    - Suppose suppose that the distribution of Y has a parameter that we will call θ for now π
    - Let the probability mass function be denoted by $f(y|\theta)$
  - Then *assuming that the data are drawn independently from Y*, then the likelihood is the product of the PMFs evaluated at the observed values $y_1, \ldots, y_n$:

$$L(\theta|y_1, \ldots y_n) = f(y_1|\theta) \times f(y_2|\theta) \times \ldots \times f(y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

    where $\prod$ is like the $\sum$ symbol, only for multiplication instead of addition.
    - It is not important that you know how to compute this.
    - You *should* know that the likelihood is based on the probabilities of each observed sample value, *assuming a particular value for the parameter in question.*



max. likelihood est.

- The MLE is the value of θ that makes $L(\theta|y_1, \ldots y_n)$ as large as it can be, given the data.
  - You can think of plugging in a bunch of values of θ, evaluating $L(\theta|y_1, \ldots y_n)$, and choosing the θ that results in the largest L
  - What we actually do is cleverer than that and involves iterative numerical algorithms
    - * The likelihood usually involves a long series of products, and many distributions are functions of exponentiated quantities (e.g., Normal, Binomial).
    - * Mathematically convenient to work with the (natural) log of the likelihood rather than the likelihood itself.
    - * Note that taking logs of numbers does not change their ordering, so if a value of θ provides the maximum value of the likelihood, it also gives the maximum log-likelihood.
    - * Furthermore, through calculus and some special computational procedures (which I won't require you to do) we can make the process of maximizing the function computationally efficient.
      - · Sometimes can solve equations to yield a formula for the MLE
      - · Other times iteratively update and improve the estimate until the computational algorithm "converges" to a final value.
  - The symbol for the MLE is distinguished from the parameter symbol by adding a "hat":
    - * $\hat{\theta}$ in the generic case
    - * $\hat{\pi}$ in the $Bin(n, \pi)$ example

taking ln of the likelihood won't change the position of max.

EXAMPLE: Binomial probability parameter π (Lecture 2 scripts.R)

Consider the $Bin(100, \pi)$ distribution again. We observe only one value of the RV W: $w = 20$. We want to find $\hat{\pi}$. We know that

$$f(w|\pi) = \frac{n!}{w!(n-w)!}\pi^w(1-\pi)^{(n-w)}$$

so

$$L(\pi|w) = f(w|\pi) = \frac{100!}{20!(80)!}\pi^{20}(1-\pi)^{80}.$$

We could derive almost the same likelihood by returning to the original binary observations $y_1, \ldots, y_{100}$, where 20 of them are 1 and the rest are 0. Recall that a binary RV has a Bernoulli distribution with PMF $f(y|\pi) = \pi^y(1-\pi)^{(1-y)}$. Then,

$$L(\pi|y_1, y_2, \ldots, y_{100}) = \prod_{i=1}^{100} \pi^{y_i}(1-\pi)^{(1-y_i)} = \pi^{\sum y_i}(1-\pi)^{(n-\sum y_i)} = \pi^{20}(1-\pi)^{80}.$$

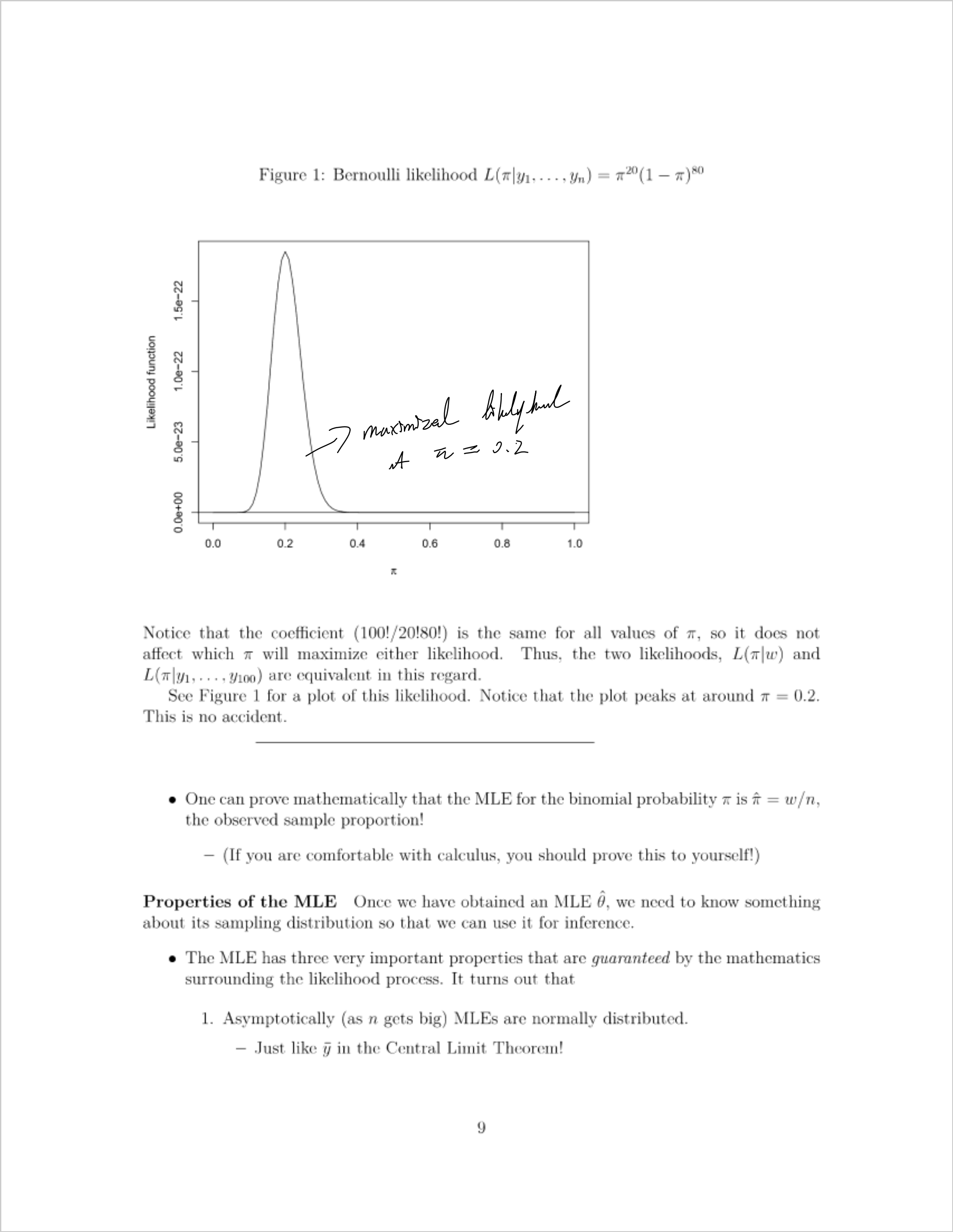


Figure 1: Bernoulli likelihood $L(\pi|y_1, \ldots, y_n) = \pi^{20}(1-\pi)^{80}$

Likelihood function
0.0e+00
5.0e−23
1.0e−22
1.5e−22
0.0
0.2
0.4
0.6
0.8
1.0
π
→ maximized likelyhood
at π = 0.2

Notice that the coefficient (100!/20!80!) is the same for all values of π, so it does not affect which π will maximize either likelihood. Thus, the two likelihoods, $L(\pi|w)$ and $L(\pi|y_1, \ldots, y_{100})$ are equivalent in this regard.

See Figure 1 for a plot of this likelihood. Notice that the plot peaks at around $\pi = 0.2$. This is no accident.

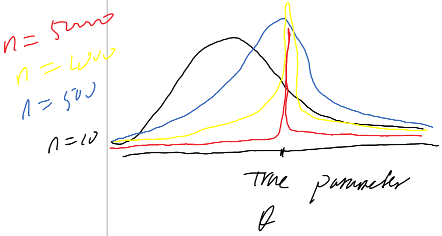
- One can prove mathematically that the MLE for the binomial probability π is $\hat{\pi} = w/n$, the observed sample proportion!
  - (If you are comfortable with calculus, you should prove this to yourself!)

Properties of the MLE Once we have obtained an MLE $\hat{\theta}$, we need to know something about its sampling distribution so that we can use it for inference.

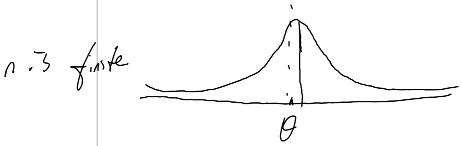
- The MLE has three very important properties that are *guaranteed* by the mathematics surrounding the likelihood process. It turns out that
  1. Asymptotically (as n gets big) MLEs are normally distributed.
     - Just like $\bar{y}$ in the Central Limit Theorem!

    - For any finite sample (i.e., *always*) the MLE has *approximately* a normal sampling distribution, and the quality of the approximation is good when the sample size is "large enough" and poor when it is "too small".

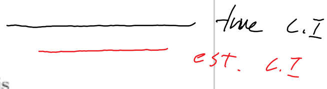
2. The MLE is CONSISTENT: as n grows, the MLE's sampling distribution closes in on the true value of the parameter
    - This is a comfort, knowing that in large samples, the MLE is estimating the "right" quantity.
    - In small samples, there is no guarantee that the MLE's sampling distribution is centered on the parameter
        - ∗ i.e., it *may* be biased in finite samples
        - ∗ In the case of a sample mean it is not, but other statistics that are MLEs may be.



3. Asymptotically (as n gets big) the MLE is the most precise (least variable) estimate you can get.
    - For estimating the middle of the Normal distribution, you could choose many different estimates, for example the mean or the median. The mean turns out to be the MLE for that case, and so it has smaller asymptotic standard error than all alternatives.

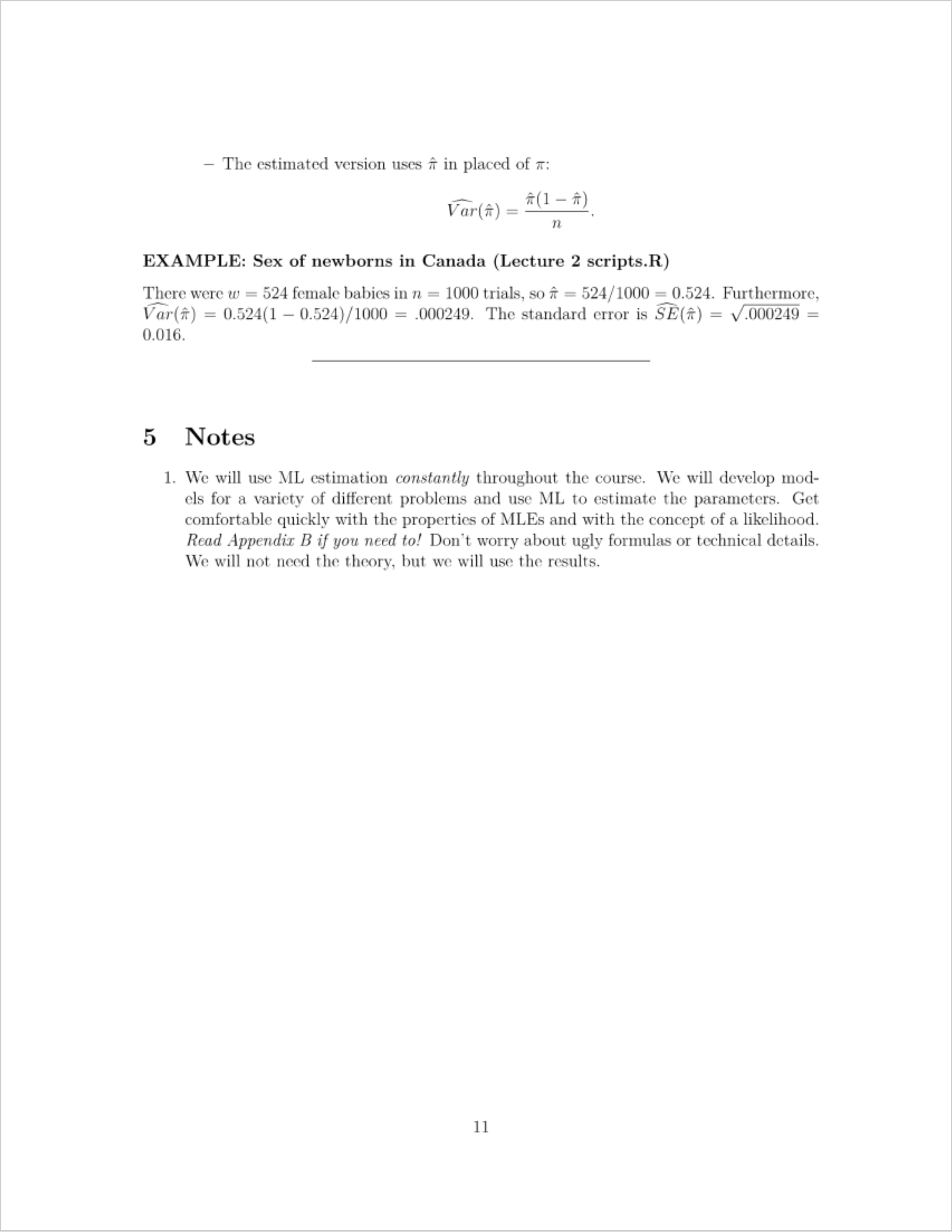


- Thus, in large samples (whatever that means), MLEs are excellent
    - In smaller samples, they may not be, but it is hard to find something else that is, so we use them anyway
    - Be cautious about drawing firm conclusions in smaller samples
- The variance of an MLE, $Var(\hat{\theta})$, is estimated using mathematical formulas described in Appendix B.3.4.
    - They are based on further manipulations of the likelihood function, so they are always available for use in conjunction with an MLE
        - ∗ I won't make you learn them!
    - The standard error of the estimate is just the square root of the variance, $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$
    - These formulas depend on the data, so estimate them using sample quantities, which puts a hat on their symbols: $\widehat{SE}(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})}$
        - ∗ Be aware that estimated standard errors often *underestimate* the actual variability in the MLE
        - ∗ The problem is worse in small samples



- In the case of the binomial probability parameter π, the variance of $\hat{\pi} = w/n$ is

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}.$$



- The estimated version uses $\hat{\pi}$ in placed of π:

$$\widehat{Var}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{n}.$$

EXAMPLE: Sex of newborns in Canada (Lecture 2 scripts.R)

There were $w = 524$ female babies in $n = 1000$ trials, so $\hat{\pi} = 524/1000 = 0.524$. Furthermore, $\widehat{Var}(\hat{\pi}) = 0.524(1-0.524)/1000 = .000249$. The standard error is $\widehat{SE}(\hat{\pi}) = \sqrt{.000249} = 0.016$.

5 Notes

1. We will use ML estimation *constantly* throughout the course. We will develop models for a variety of different problems and use ML to estimate the parameters. Get comfortable quickly with the properties of MLEs and with the concept of a likelihood. *Read Appendix B if you need to!* Don't worry about ugly formulas or technical details. We will not need the theory, but we will use the results.