Contents

3 Parameter and Interval Estimation

In the previous chapter, we discussed random variables and developed the notion of a probability distribution, and then established some fundamental results such as the central limit theorem that give strong and useful information about the statistical properties of a sample drawn from a fixed, known distribution.

Our goal in this chapter is, in some sense, to invert this analysis: starting instead with data obtained by sampling a distribution or probability model with certain unknown parameters, we would like to extract information about the most reasonable values for these parameters given the observed data. We begin by discussing pointwise parameter estimates and estimators, analyzing various properties that we would like these estimators to have such as unbiasedness and efficiency, and finally establish the optimality of several estimators that arise from the basic distributions we have encountered such as the normal and binomial distributions.

We then broaden our focus to interval estimation: rather than merely finding the best estimate of a single value, we seek to find this best estimate along with a measurement of its expected precision. We treat in scrupulous detail several important cases for constructing such confidence intervals, and close with some applications of these ideas to polling data.

3.1 Parameter Estimation

- To motivate our formal development of estimation methods, we first outline a few scenarios in which we would like to use parameter estimation. For reasons of consistency, we will always call our unknown parameter θ.
    - As a first example, suppose that an unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH. We would like to know what the most reasonable estimate for θ is, given these results.
    - In this case, it seems reasonable to say that since 8 of the flips are tails and 2 of the flips are heads, the most reasonable estimate for θ would be $2/10 = 0.2$. It seems far more likely that we would obtain the results above with a coin that has a $1/5$ chance of landing heads (since then the expected number of heads in 10 flips is 2, exactly what we observed) than, say, if the coin had a $1/2$ chance of landing heads (since then the expected number of heads would be 5, far more than we observed).

- As a second example, suppose that we expect the number of calls received by an emergency help line at night should have a Poisson distribution with parameter $\lambda = \theta$. If the number of calls received in consecutive hours is 4, 2, 0, 3, and 4 respectively, what is the most reasonable estimate for the associated parameter θ?
- As a third example, suppose that we sample a continuous random variable that is known to be exponentially distributed with some parameter $\lambda = \theta$. If the values obtained from the samples are 1.25, 0.02, 0.18, and 0.63, what is the most reasonable estimate for the associated parameter θ?

- In these last two examples, it is less obvious how we might go about estimating the most reasonable value for the parameter.
  - One possible approach for the second example is as follows: in terms of θ, we compute the probability of obtaining the sampling data we received. Then we search among the possible values of θ for the one that makes our observed outcomes most likely to have occurred.
  - We can take a similar approach for the third example, provided we use the probability density function in place of the actual probabilities of obtaining the observed values (since they will always be zero).

3.1.1 Maximum Likelihood Estimates

- As motivated above, the quantity we would like to maximize is the likelihood of obtaining the observed data as a function of the parameter θ, which we can define as follows:

- Definition: Suppose the values $x_1, x_2, \ldots, x_n$ are observed by sampling a discrete or continuous random variable X with probability density function $f_X(x;\theta)$ that depends upon an unknown parameter θ. Then the likelihood function $L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = f_X(x_1;\theta) \cdot f_X(x_2;\theta) \cdot \cdots \cdot f_X(x_n;\theta)$ represents the probability associated to the observed values x_i.
  - In the situation where X is a discrete random variable, then under the assumption that all of the samples are independent, the product $f_X(x_1;\theta) \cdot f_X(x_2;\theta) \cdot \cdots \cdot f_X(x_n;\theta)$ represents the probability of obtaining the outcomes $x_1, x_2, \ldots, x_n$ from sampling X a total of n times in a row.
  - In the situation where X is a continuous random variable, the product represents the probability density of obtaining that sequence of outcomes.
  - In either scenario, we think of the likelihood function $L(\theta)$ as measuring the overall probability that we would obtain the observed data by sampling the distribution having parameter θ.
  - Example: If an unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH, the likelihood function is $L(\theta) = (1-\theta)^6 \cdot \theta \cdot (1-\theta)^2 \cdot \theta = \theta^2(1-\theta)^8$.
  - Example: If a Poisson distribution with parameter $\lambda = \theta$ is sampled five times and the results are 4, 2, 0, 3, 4, the likelihood function is $L(\theta) = \left[\frac{\theta^4 e^{-\theta}}{4!}\right] \cdot \left[\frac{\theta^2 e^{-\theta}}{2!}\right] \cdot \left[\frac{\theta^0 e^{-\theta}}{0!}\right] \cdot \left[\frac{\theta^3 e^{-\theta}}{3!}\right] \cdot \left[\frac{\theta^4 e^{-\theta}}{4!}\right] = \frac{\theta^{13} e^{-5\theta}}{6912}$.
  - Example: If an exponential distribution with parameter $\lambda = \theta$ is sampled five times with results 1.25, 0.02, 0.18, 0.63, the likelihood function is $L(\theta) = [\theta e^{-1.25\theta}] \cdot [\theta e^{-0.02\theta}] \cdot [\theta e^{-0.18\theta}] \cdot [\theta e^{-0.63\theta}] = \theta^4 e^{-2.08\theta}$.

- Our approach now is to compute the value of the unknown parameter that maximizes the likelihood of obtaining the observed data; this is known as the method of maximum likelihood. Here is a more explicit description of the method:

- Method (Maximum Likelihood): Suppose the values $x_1, x_2, \ldots, x_n$ are observed by sampling a random variable X with probability density function $f_X(x;\theta)$ that depends upon an unknown parameter θ. Then a maximum likelihood estimate (MLE) for θ, often written as $\hat{\theta}$ or θ_e, is a value of θ that maximizes the likelihood function $L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)$.
  - Before giving an example, we will observe that, at least in principle, there could be more than one value of θ maximizing the function $L(\theta)$. In practice, there is usually a unique maximum, which we refer to as *the* maximum likelihood estimate of θ.

- In the event that the function $f_X(x_i;\theta)$ is a differentiable function of θ (which is usually the case) then $L(\theta)$ will also be a differentiable function of θ: thus, by the usual principle from calculus, any maximum likelihood estimate will be a global maximum of L hence be a root of the derivative $L'(\theta)$.
- Since $L(\theta)$ is a product, to compute the roots of its derivative it is much easier instead to use logarithmic differentation, which amounts to computing the roots of the derivative of its logarithm $\ln L(\theta) = \sum_{i=1}^{n} \ln[f_X(x_i;\theta)]$, which is called the log-likelihood.

- Example: An unfair coin with an unknown probability θ of coming up heads is flipped 10 times, and the results are TTTTT THTTH. Find the maximum likelihood estimate for θ.
  - Above, we found the likelihood function $L(\theta) = (1-\theta)^6 \cdot \theta \cdot (1-\theta)^2 \cdot \theta = \theta^2(1-\theta)^8$.
  - The log-likelihood is $\ln L(\theta) = 2\ln(\theta) + 8\ln(1-\theta)$ with derivative $\dfrac{d}{d\theta}[\ln L(\theta)] = \dfrac{2}{\theta} - \dfrac{8}{1-\theta}$.
  - Setting the derivative equal to zero yields $\dfrac{2}{\theta} - \dfrac{8}{1-\theta} = 0$ so that $2(1-\theta) = 8\theta$, whence $\theta = \boxed{1/5}$.
  - Remark: Note that this result agrees with our intuitive argument earlier that the most reasonable value of θ is the actual proportion of heads obtained in the sample.
  - Remark: We could, of course, work instead with the derivative of the original likelihood function $L'(\theta) = 2\theta(1-\theta)^8 - 8\theta^2(1-\theta)^7 = \theta(1-\theta)^7(2(1-\theta) - 8\theta)$. Setting $L'(\theta) = 0$ and solving yields $\theta = 0, 1, 1/5$. Notice that although $\theta = 0$ and $\theta = 1$ are roots of $L'(\theta) = 0$, and hence are critical numbers for $L(\theta)$, they are in fact local minima, whereas $\theta = 1/5$ is a local maximum. We implicitly ignored the two values $\theta = 0$ and $\theta = 1$ when analyzing the log-likelihood because these make $\frac{d}{d\theta}\ln L(\theta)$ undefined rather than zero. In principle, we should always check that the candidate value actually does yield the *maximum* likelihood, but we will omit such verifications when there is only one possible candidate.

- Example: A Poisson distribution with parameter $\lambda = \theta$ representing the number of calls to an emergency help line is sampled five times and the results are 4, 2, 0, 3, 4. Find the maximum likelihood estimate for θ.
  - Above, we got the likelihood function $L(\theta) = \left[\dfrac{\theta^4 e^{-\theta}}{4!}\right] \cdot \left[\dfrac{\theta^2 e^{-\theta}}{2!}\right] \cdot \left[\dfrac{\theta^0 e^{-\theta}}{0!}\right] \cdot \left[\dfrac{\theta^3 e^{-\theta}}{3!}\right] \cdot \left[\dfrac{\theta^4 e^{-\theta}}{4!}\right] = \dfrac{\theta^{13} e^{-5\theta}}{6912}$.
  - Then the log-likelihood is $\ln L(\theta) = 13\ln(\theta) - 5\theta - \ln(6912)$, with derivative $\dfrac{d}{d\theta}[\ln L(\theta)] = \dfrac{13}{\theta} - 5$. This is equal to zero for $\theta = \boxed{13/5}$, so this value is our maximum likelihood estimate.
  - Remark: Notice that this value $\theta = 13/5$ represents the average number of calls to the help line per hour in the data sample. Intuitively, because the parameter λ for the Poisson distribution represents the expected value (which in this case represents the average number of calls per hour), it is also very reasonable to find that the sample average is a good estimate λ.

- Example: An exponential distribution with parameter θ is sampled five times and the results are 1.25, 0.02, 0.18, 0.63. Find the maximum likelihood estimate for θ.
  - Above, we computed the likelihood function $L(\theta) = [\theta e^{-1.25\theta}] \cdot [\theta e^{-0.02\theta}] \cdot [\theta e^{-0.18\theta}] \cdot [\theta e^{-0.63\theta}] = \theta^4 e^{-2.08\theta}$.
  - Then the log-likelihood is $\ln L(\theta) = 4\ln\theta - 2.08\theta$, with derivative $\dfrac{d}{d\theta}[\ln L(\theta)] = \dfrac{4}{\theta} - 2.08$. This is equal to zero for $\theta = \boxed{4/2.08} \approx 1.9231$, so this value is our maximum likelihood estimate.
  - Remark: If we again observe that the expected value of the exponential distribution is $1/\theta$, if we set this equal to the observed expected value 2.08/4, we obtain the maximum likelihood estimate for θ.

- Example: A normal distribution with mean 0 and standard deviation θ is sampled four times and the results are 2.08, 0.34, -2.65, and 2.28. Find the maximum likelihood estimate for θ.
  - The probability density function for this normal distribution is $f_X(x;\theta) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-x^2/(2\theta^2)}$.
  - Thus, $\ln f_X(x;\theta) = -\ln(\sqrt{2\pi}) - \ln(\theta) - \dfrac{x^2}{2\theta^2}$. Now sum the appropriate values to obtain the log-likelihood: $\ln L(\theta) = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{2.08^2}{2\theta^2} - \dfrac{0.34^2}{2\theta^2} - \dfrac{(-2.65)^2}{2\theta^2} - \dfrac{2.28^2}{2\theta^2} = 4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{16.6629}{2\theta^2}$.

- The derivative is then $\frac{d}{d\theta}[\ln L(\theta)] = -\frac{4}{\theta} + \frac{16.6629}{\theta^3}$, which is zero for $\theta = \pm\sqrt{\frac{16.6629}{4}} \approx \pm 2.0410$.
- Since the standard deviation is nonnegative, the maximum likelihood estimate is $\boxed{\sqrt{\frac{16.6629}{4}}} \approx 2.0410$.

- Example: A normal distribution with mean θ and standard deviation θ is sampled four times and the results are 14, -3, 12, and 8. Find the maximum likelihood estimate for θ.
  - The probability density function for this normal distribution is $f_X(x;\theta) = \frac{1}{\theta\sqrt{2\pi}} e^{-(x-\theta)^2/(2\theta^2)}$.
  - Thus, $\ln f_X(x;\theta) = -\ln(\sqrt{2\pi}) - \ln(\theta) - \frac{(x-\theta)^2}{2\theta^2}$. We then sum the appropriate values to obtain the log-likelihood: $\ln L(\theta) = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \frac{(14-\theta)^2}{2\theta^2} - \frac{(-3-\theta)^2}{2\theta^2} - \frac{(12-\theta)^2}{2\theta^2} - \frac{(8-\theta)^2}{2\theta^2} = 4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \frac{413 - 62\theta + 4\theta^2}{2\theta^2}$.
  - The derivative is then $\frac{d}{d\theta}[\ln L(\theta)] = -\frac{4}{\theta} + \frac{413}{\theta^3} - \frac{31}{\theta^2}$. Setting this equal to zero and clearing denominators yields $-4\theta^2 - 31\theta + 413 = 0$ which has roots $\theta = -\frac{59}{4}, 7$.
  - Since the standard deviation is always nonnegative, the maximum likelihood estimate is $\theta = \boxed{7}$.

- In some cases, after taking the derivative of the log-likelihood we may be left with an equation that cannot be solved analytically for θ (unlike the examples above, where we could always solve explicitly for θ). In such cases, we must resort to numerical approximation procedures, such as Newton's method, to find the desired root.

- Example: The continuous random variable with probability density function $f_X(x;\theta) = \frac{2(\theta - x)}{\theta^2}$ for $0 \le x \le \theta$ is sampled five times, and the results are 1.31, 0.83, 1.19, 0.20, and 0.06. Find the maximum likelihood estimate for θ.
  - We have $\ln f_X(x;\theta) = \ln 2 + \ln(\theta - x) - 2\ln(\theta)$, so now we sum the appropriate values to obtain the log-likelihood: $\ln L(\theta) = 5\ln 2 + \ln(\theta - 1.31) + \ln(\theta - 0.83) + \ln(\theta - 1.19) + \ln(\theta - 0.2) + \ln(\theta - 0.06) - 10\ln(\theta)$.
  - The derivative is then $\frac{d}{d\theta}[\ln L(\theta)] = \frac{1}{\theta - 1.31} + \frac{1}{\theta - 0.83} + \frac{1}{\theta - 1.19} + \frac{1}{\theta - 0.20} + \frac{1}{\theta - 0.06} - \frac{10}{\theta}$.
  - Setting the derivative equal to zero and clearing denominators yields a polynomial equation of degree 5 in θ, whose roots cannot be easily evaluated[1].
  - Using Newton's method or another numerical approximation technique, we can find approximations to the roots as $\theta \approx 0.0677$, 0.2308, 0.9113, 1.2460, and 2.3307.
  - Since one of the observed values was 1.31, and the density function is only nonzero when $0 \le x \le \theta$, we must have $\theta \ge 1.31$. Therefore, the maximum likelihood estimate can only be the largest root, $\theta \approx \boxed{2.3307}$.

- It is also possible to perform maximum likelihood estimates for more than one unknown parameter simultaneously.
  - The idea is the same as in the single-parameter case: we write down the likelihood function and then attempt to maximize it.
  - For a differentiable function of several variables, any local maximum must occur at a point where all partial derivatives of the function are zero.
  - As in the one-parameter case, since the likelihood function is a product, we usually work instead with the log-likelihood function.

[1] In fact, it is a theorem from abstract algebra (Abel's theorem) that there does not exist any elementary formula in radicals for the solution to a general degree-5 polynomial.

- What this means is that we may find a multi-parameter maximum likelihood estimate by setting all of the partial derivatives of the log-likelihood function equal to zero, and then solving the resulting system of equations for the unknown parameters.

- Example: A normal distribution with unknown mean μ and standard deviation θ is sampled three times and the results are 1, 5, and -3. Find the maximum likelihood estimate for (μ, θ).
  - The probability density function for this normal distribution is $f_X(x;\theta,\mu) = \dfrac{1}{\theta\sqrt{2\pi}} e^{-(x-\mu)^2/(2\theta^2)}$.
  - Thus, the log-likelihood function is $\ln L(\theta,\mu) = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{(1-\mu)^2}{2\theta^2} - \dfrac{(5-\mu)^2}{2\theta^2} - \dfrac{(-3-\mu)^2}{2\theta^2} = -4\ln(\sqrt{2\pi}) - 4\ln(\theta) - \dfrac{35-6\mu+3\mu^2}{2\theta^2}$.
  - The partial derivatives are $\dfrac{\partial}{\partial\mu}[\ln L(\theta,\mu)] = \dfrac{6-6\mu}{2\theta^2}$ and $\dfrac{\partial}{\partial\theta}[\ln L(\theta,\mu)] = -\dfrac{4}{\theta} + \dfrac{35-6\mu+3\mu^2}{\theta^3}$.
  - Setting the partial derivatives equal to zero and solving yields, respectively, $6-6\mu=0$ so that $\mu=1$, and $\theta^2 = \dfrac{35-6\mu+3\mu^2}{4} = 8$ so that $\theta = \pm\sqrt{8}$.
  - Since the standard deviation is positive, we obtain the maximum likelihood estimates $\boxed{\mu = 1}$ and $\boxed{\theta = \sqrt{8}}$.

- With more complicated functions of several parameters, the resulting system of equations can be very difficult to solve, even with numerical methods.
  - For this reason, certain other methods are used in lieu of maximum likelihood estimates.
  - One such method is known as the method of moments. This method involves computing the so-called moments $E(X^k)$ for integers $k = 1, 2, \ldots, n$ where n is the total number of unknown parameters, and then setting them equal to the corresponding moments of the sample data. The resulting system of equations is often much easier to solve than the system arising from a maximum likelihood estimate.
  - In many cases, the estimates yielded by the method of moments are a good approximation to those arising from maximum likelihood estimates (and for many common distributions, they are often the same), and can be used as a starting point for approximation methods.
  - For the one-parameter case, the method of moments is the same as requiring that the estimate's expected value agrees with the sample's expected value, while in the two-parameter case, since $\text{var}(X) = E(X^2) - E(X)^2$, it is the same as requiring that the estimate's expected value and variance agree with the sample's expected value and variance.

3.1.2 Biased and Unbiased Estimators

- Instead of performing a maximum likelihood estimate for each set of sample data, we can instead try to write down a general formula for the estimate in terms of the data values we observe: such a function is an estimator for our parameter of interest.

- Example: A Poisson distribution with parameter θ is sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimator $\hat{\theta}(x_1, \ldots, x_n)$ for θ in terms of $x_1, x_2, \ldots, x_n$.
  - Since $f_X(x;\theta) = \dfrac{\theta^x e^{-\theta}}{x!}$, the log-likelihood function is $\ln L(\theta) = (x_1 + x_2 + \cdots + x_n)\ln(\theta) - n\theta - \ln(x_1! x_2! \cdots x_n!)$.
  - Thus, $\dfrac{d}{d\theta}[\ln L(\theta)] = \dfrac{x_1 + x_2 + \cdots + x_n}{\theta} - n$, which is equal to zero for $\hat{\theta} = \boxed{\dfrac{x_1 + x_2 + \cdots + x_n}{n}}$.

- Example: A normal distribution with unknown mean μ and standard deviation σ is sampled sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimators $\hat{\mu}(x_1, \ldots, x_n)$ and $\hat{\sigma}(x_1, \ldots, x_n)$.

- Since $f_X(x;\mu,\sigma) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$, the log-likelihood function is $\ln L(\mu,\sigma) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \dfrac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{2\sigma^2}$.
- The partial derivatives are $\dfrac{\partial}{\partial\mu}[\ln L(\mu,\sigma)] = \dfrac{(x_1-\mu) + (x_2-\mu) + \cdots + (x_n-\mu)}{\sigma^2}$ and $\dfrac{\partial}{\partial\sigma}[\ln L(\mu,\sigma)] = -\dfrac{n}{\sigma} + \dfrac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{\theta^3}$.
- Setting the partial derivatives equal to zero and solving yields, respectively, $\mu = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ and $\sigma^2 = \dfrac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{n} = \dfrac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) - \left(\dfrac{x_1 + x_2 + \cdots + x_n}{n}\right)^2$.
- Notice here that the resulting values of μ and σ are simply the mean and standard deviation of the outcome set $\{x_1, x_2, \ldots, x_n\}$.

- Example: A uniform distribution on $[0,\theta]$ is sampled n times yielding outcomes $x_1, x_2, \ldots, x_n$. Find the maximum likelihood estimator $\hat{\theta}$ for θ.
  - The probability density function for this uniform distribution is $f_X(x;\theta) = \begin{cases} 1/\theta & \text{for } 0 \le x \le \theta \\ 0 & \text{for other } x \end{cases}$.
  - Therefore, the likelihood function is $L(\theta) = \begin{cases} 1/\theta^n & \text{if } x_1, x_2, \ldots x_n \le \theta \\ 0 & \text{otherwise} \end{cases}$.
  - Since $1/\theta^n$ decreases with increasing θ, we can see that the maximum value will occur for the smallest possible θ for which the first condition is satisfied, which is $\theta = \max(x_1, x_2, \ldots, x_n)$.
  - Thus, the maximum likelihood estimator is $\hat{\theta} = \boxed{\max(x_1, x_2, \ldots, x_n)}$.
  - Remark: The discrete analogue of this particular problem is known as the German tank problem: during World War II, British intelligence was able to capture numerous components from German tanks, each of which was stamped with its manufacturing number. The labels were thus effectively drawn at random from $[0,\theta]$ where θ was the total number of German tanks. For obvious reasons, it was of substantial military interest to estimate as precisely as possible the total number θ of enemy tanks; the (quite surprising) result of this calculation shows the largest part number that was observed is actually a good estimate. As a historical matter, the projections obtained by the statisticians analyzing this problem were far more accurate than those obtained by other methods!

- In general, there are many possible estimators for any given parameter, and it is not always clear which one we should use.
  - For example, in the German tank problem discussed above, it does not seem entirely reasonable that the "best estimate" $\hat{\theta}$ for the number of enemy tanks is simply the largest number observed: since this estimate is always the lowest feasible number of tanks that is consistent with the observed data, and there is a nontrivial chance that the actual number of tanks is larger than $\hat{\theta}$. We would expect that the maximum likelihood estimate should, in general, tend to underestimate the actual correct value of θ.

- Intuitively, we might prefer to search for an estimator that tends to have the smallest systematic error.
  - The most basic possible requirement is to ask that the estimator not have any "bias", on average, away from the expected value of the parameter.
  - To study this more precisely, we will shift our emphasis to view estimators as random variables on the space of possible input data.
  - When we think of the estimator as a random variable, we can equivalently phrase this requirement for a lack of bias in terms of the expected value:

- Definition: We say that an estimator $\hat{\theta}(x_1, x_2, \ldots, x_n)$ for a set of observations $x_1, \ldots, x_n$ drawn by randomly sampling a random variable X with probability density function $f_X(x;\theta)$ is unbiased if $E(\hat{\theta}) = \theta$ for all θ.

- More verbosely, this means that if we fix θ and then average over all possible samples $x_1, \ldots, x_n$ of X for a fixed value of θ, then $\hat{\theta}$ is unbiased when the expected value of the estimator $\hat{\theta}$ is equal to the true value of the parameter θ.

- Example: Show that the maximum likelihood estimate $\hat{\mu} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ for sampling the normal distribution with mean μ and fixed standard deviation σ is unbiased.
  - Note that by properties of expected value, we have $E(\hat{\mu}) = \dfrac{1}{n}[E(x_1) + E(x_2) + \cdots + E(x_n)]$.
  - Furthermore, we have $E(x_i) = \mu$ because each x_i is sampled randomly from a distribution with mean μ.
  - Thus, we have $E(\hat{\mu}) = \dfrac{1}{n}[n\mu] = \mu$, and so μ is unbiased.
  - Remark: Note, more generally, that the estimator $\hat{\mu} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$ for any choice of constants a_i such that $\sum_i a_i = 1$ will be unbiased, by the same calculation. Thus, for example, the estimator $\frac{1}{3}x_1 + \frac{2}{3}x_2$ is also unbiased.

- Example: Show that the maximum likelihood estimate for the variance $\hat{\sigma}^2 = \dfrac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) - \left[\dfrac{x_1 + x_2 + \cdots + x_n}{n}\right]^2$ from sampling the normal distribution with unknown mean μ and standard deviation σ is biased.
  - Recall that $\mu = E(x_i)$ and $\sigma^2 = E(x_i^2) - E(x_i)^2$, so $E(x_i^2) = \sigma^2 + \mu^2$.
  - Furthermore, because variance is additive for independent random variables, we see that $\mathrm{var}(x_1 + x_2 + \cdots + x_n) = \mathrm{var}(x_1) + \mathrm{var}(x_2) + \cdots + \mathrm{var}(x_n) = n\sigma^2$.
  - Since $\mathrm{var}(S) = E(S^2) - E(S)^2$, applying this for $S = x_1 + x_2 + \cdots + x_n$ yields $E(S^2) = \mathrm{var}(S) + E(S)^2 = n\sigma^2 + n^2\mu^2$.
  - Then by properties of expected value, we have $E(\hat{\sigma}^2) = \dfrac{1}{n}E(x_1^2 + x_2^2 + \cdots + x_n^2) - \dfrac{1}{n^2}E(x_1 + x_2 + \cdots + x_n)^2 = \dfrac{1}{n} \cdot n(\sigma^2 + \mu^2) - \dfrac{1}{n^2}(n\sigma^2 + n^2\mu^2) = \dfrac{n-1}{n}\sigma^2$.
  - The expected value is not equal to σ^2 because of the factor of $\dfrac{n-1}{n}$, so this estimator is biased.

- In the example above, we can construct an unbiased estimator of σ^2 by scaling $\hat{\sigma}^2$ by $\dfrac{n}{n-1}$.
  - This new estimator $S^2 = \dfrac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$, where $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ is the sample average, is called the sample variance.
  - The square root $S = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2}$ of the sample variance is called the sample standard deviation.
  - Despite the fact that $E(S) \neq \sigma^2$, S is quite commonly used as an estimator for σ because the estimate of σ^2 by S^2 is unbiased.
  - We can give some intuitive motivation for why the factor $\dfrac{1}{n-1}$ appears in the sample standard deviation[3]: imagine trying to estimate the variance in the sizes of members of a newly-discovered species. If only one member of the species is found, there is no way to give a reasonable estimate in the variation, and so any plausible value for the variance is reasonable (thus the formula should be undefined or infinite when $n = 1$). On the other hand, if two members of the species are found, then the difference in their sizes gives a plausible range of variation (thus the formula should be well-defined when $n = 2$).

[2] It is not entirely trivial to compute the expected value of S itself, but a formula does exist in terms of the gamma function: $E(S) = \sigma \cdot \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$, where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$. In particular, the constant factor on the right-hand side is approximately $1 - \frac{1}{4n}$ for large n.

[3] The use of $n-1$ in place of n in the denominator of the sample variance is known as Bessel's correction. Roughly speaking, the correction is required because measuring the variance of the sample relative to the sample mean (rather than relative to the unknown true mean μ) will always lower the estimated variance, so, as we calculated, a correction is needed to unbias the estimate.

- Example: Show that the maximum likelihood estimator $\hat{\theta} = \max(x_1, x_2, \ldots, x_n)$ from sampling the uniform distribution on $[0, \theta]$ is biased.
  - Here, we need to compute the expected value of $\hat{\theta}$, which requires us to find the underlying probability distribution.
  - Observe that, for any $0 \leq x \leq \theta$, we have $P(\hat{\theta} \leq x) = (x/\theta)^n$ because $\hat{\theta} \leq k$ occurs precisely when all of the values $x_1, x_2, \ldots, x_n$ lie in the interval $[0, x]$, which occurs with probability $(x/\theta)^n$.
  - This means that the cumulative distribution function for $\hat{\theta}$ is $g_{\hat{\theta}}(x) = (x/\theta)^n$ for $0 \leq x \leq \theta$, and so its probability distribution function is the derivative $g'_{\hat{\theta}}(x) = nx^{n-1}/\theta^n$.
  - Then we may compute $E(\hat{\theta}) = \int_0^\theta x g'_{\hat{\theta}}(x)\, dx = \int_0^\theta nx^n/\theta^n\, dx = \dfrac{n}{n+1}\theta$. Since this is not equal to θ, we see that $\hat{\theta}$ is biased, as claimed.
  - Remark: Like with the sample variance, we can rescale the maximum likelihood estimate to obtain an unbiased estimator of θ, namely, $\hat{\theta} = \dfrac{n+1}{n}\max(x_1, x_2, \ldots, x_n)$.

- Example: Show that the estimator $\hat{\theta} = \dfrac{2}{n}(x_1 + x_2 + \cdots + x_n)$ from sampling the uniform distribution on $[0, \theta]$ is unbiased.
  - Since each x_i is drawn from the uniform distribution on $[0, \theta]$, we have $E(x_i) = \int_0^\theta x \cdot \frac{1}{\theta} dx = \dfrac{\theta}{2}$.
  - Then, by properties of expected value, we have $E(\hat{\theta}) = \dfrac{2}{n}\left(E[x_1] + E[x_2] + \cdots + E[x_n]\right) = \dfrac{2}{n} \cdot n \cdot \dfrac{\theta}{2} = \theta$.
  - Therefore, $\hat{\theta}$ is unbiased, as claimed.

3.1.3 Efficiency of Estimators

- As we have already remarked, for any given parameter estimation problem, there are many different possible choices for estimators.
  - One desirable quality for an estimator is that it be unbiased. However, this requirement alone does not impose a substantial condition, since (as we have seen) there can exist several different unbiased estimators for a given parameter.
  - For example, we have shown that for estimating the parameter θ, given a random sample $x_1, x_2, \ldots, x_n$ from the uniform distribution on $[0, \theta]$, both of the estimators $\hat{\theta}_1 = \dfrac{n+1}{n}\max(x_1, x_2, \ldots, x_n)$ and $\hat{\theta}_2 = \dfrac{2}{n}(x_1 + x_2 + \cdots + x_n)$ are unbiased.
  - Likewise, we have also shown that, given a random sample x_1, x_2 from the normal distribution with mean θ and standard deviation σ, the estimators $\hat{\theta}_1 = \frac{1}{2}(x_1 + x_2)$ and $\hat{\theta}_2 = \frac{1}{3}(x_1 + 2x_2)$ are also both unbiased.
  - We would now like to know if there is a meaningful way to say one of these unbiased estimators is better than the other.
  - In the abstract, it seems reasonable to say that an estimator with a smaller variance is better than one with a larger variance, since a smaller variance would indicate that the value of the estimator stays closer to the "true" parameter value more often.

- Definition: If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for the parameter θ, we say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$.

- Example: Given a random sample x_1, x_2 from the normal distribution with mean θ and standard deviation σ, which of $\hat{\theta}_1 = \frac{1}{2}(x_1 + x_2)$ and $\hat{\theta}_2 = \frac{1}{3}(x_1 + 2x_2)$ is a more efficient estimator for θ?
  - Note that because x_1 and x_2 are independent, their variances are additive. We can then compute $\text{var}(\hat{\theta}_1) = \text{var}(\frac{1}{2}x_1 + \frac{1}{2}x_2) = \text{var}(\frac{1}{2}x_1) + \text{var}(\frac{1}{2}x_2) = \frac{1}{4}\text{var}(x_1) + \frac{1}{4}\text{var}(x_2) = \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{1}{2}\sigma^2$.
  - Also, $\text{var}(\hat{\theta}_2) = \text{var}(\frac{1}{3}x_1 + \frac{2}{3}x_2) = \text{var}(\frac{1}{3}x_1) + \text{var}(\frac{2}{3}x_2) = \frac{1}{9}\text{var}(x_1) + \frac{4}{9}\text{var}(x_2) = \frac{1}{9}\sigma^2 + \frac{4}{9}\sigma^2 = \frac{5}{9}\sigma^2$.

- Since $\frac{1}{2}\sigma^2 < \frac{5}{9}\sigma^2$, we see that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.
- More generally, we could ask: for an arbitrary parameter a, which estimator $\hat{\theta}_a = ax_1 + (1-a)x_2$ is the most efficient?
- In the same way as above, we can compute $\text{var}(\hat{\theta}_a) = [a^2+(1-a)^2]\sigma^2 = (2a^2-2a+1)\sigma^2$. By calculus (the derivative is $4a-2$ which is zero for $a = 1/2$) or by completing the square ($2a^2-2a+1 = 2(a-1/2)^2+1/2$) we can see that the minimum occurs when $a = 1/2$.
- Remark: Intuitively, this last calculation should make sense, because if we put more weight on one observation, its variation will tend to dominate the calculation. In the extreme situation of taking $\hat{\theta}_3 = x_2$ (which corresponds to $a = 0$), for example, we see that the variance is simply σ^2, which is much larger than the variance arising from the average. This is quite reasonable, since the average $\frac{1}{2}(x_1 + x_2)$ uses a bigger sample and thus captures more information than just using a single observation.

- Example: Given a random sample $x_1, x_2, \ldots, x_n$ from the uniform distribution on $[0, \theta]$, which of $\hat{\theta}_1 = \dfrac{n+1}{n}\max(x_1, \ldots, x_n)$ and $\hat{\theta}_2 = \dfrac{2}{n}(x_1 + x_2 + \cdots + x_n)$ is a more efficient estimator for θ?
  - To compute the variance of $\hat{\theta}_1$, first recall that we calculated the probability density function for $\max(x_1, \ldots, x_n)$ as $g(x) = nx^{n-1}/\theta^n$ for $0 \le x \le \theta$.
  - Then $E[\max(x_1, \ldots, x_n)^2] = \int_0^\theta x^2 \cdot nx^{n-1}/\theta^n \, dx = \dfrac{n}{n+2}\theta^2$, and so since $E[\max(x_1, \ldots, x_n)] = \dfrac{n}{n+1}\theta$ as we calculated previously, we see that $\text{var}[\max(x_1, \ldots, x_n)] = \dfrac{n}{n+2}\theta^2 - \left[\dfrac{n}{n+1}\theta\right]^2 = \dfrac{n}{(n+2)(n+1)^2}\theta^2$.
  - Therefore, $\text{var}(\hat{\theta}_1) = \left[\dfrac{n+1}{n}\right]^2 \text{var}[\max(x_1, \ldots, x_n)] = \dfrac{1}{n(n+2)}\theta^2$.
  - For $\hat{\theta}_2$, since the x_i are independent their variances are additive.
  - Since $\text{var}(x_i) = \dfrac{\theta^2}{12}$, we see that $\text{var}(\hat{\theta}_2) = \text{var}(\frac{2}{n}x_1) + \cdots + \text{var}(\frac{2}{n}x_n) = n \cdot \dfrac{4}{n^2} \cdot \dfrac{\theta^2}{12} = \dfrac{1}{3n}\theta^2$.
  - For $n = 1$ these variances are the same (this is unsurprising because when $n = 1$ the estimators themselves are the same!), but for $n > 1$ we see that the variance of $\hat{\theta}_1$ is smaller, so $\hat{\theta}_1$ is more efficient.

- Example: Suppose x and y are respectively drawn from two independent normal distributions X and Y with the same unknown mean $E(X) = E(Y) = \theta$ but different known variances $\text{var}(X) = \sigma^2$ and $\text{var}(Y) = 2\sigma^2$. Show that for any parameter $0 \le a \le 1$ the estimator $\hat{\theta}_a = ax + (1-a)y$ is unbiased, and find the value of a yielding the most efficient estimator.
  - By the linearity of expected value, we have $E(\hat{\theta}_a) = aE(x) + (1-a)E(y) = a\theta + (1-a)\theta = \theta$. Thus, $\hat{\theta}_a$ is unbiased for each value of a.
  - For the efficiency calculation, since x and y are independent we have $\text{var}(\hat{\theta}_a) = \text{var}[ax] + \text{var}[(1-a)y] = a^2\text{var}(x) + (1-a)^2\text{var}(y) = a^2\sigma^2 + (1-a)^2 \cdot 2\sigma^2 = (3a^2 - 4a + 2)\sigma^2$.
  - By using calculus (the derivative is $6a-4$ which is zero for $a = 2/3$) or completing the square ($3a^2-4a+2 = 3(a-2/3)^2 + 2/3$), we see that the minimum variance occurs for $a = 2/3$, so this value yields the most efficient estimator.
  - Remark: Intuitively, we should expect that weighting more closely towards x will yield a better estimate, because x has less variance than y does, so it is more likely to be closer to the true mean. Nonetheless, including y does provide some additional information, so the weighted average should not shift too far over to x.

- So far we have only discussed the relative efficiency of unbiased estimators. But since the variance of any estimator is always bounded below (since it is by definition nonnegative), it is quite reasonable to ask whether, for a fixed estimation problem, there might be an optimal unbiased estimator: namely, one of minimal variance.
  - This question turns out to be quite subtle, because we are not guaranteed that such an estimator necessarily exists.

- As a simple illustration, it could be the case that the possible variances of unbiased estimators form an open interval of the form (a, ∞) for some $a \geq 0$: then there would be estimators whose variances approach the value a arbitrarily closely, but there is none that actually achieves the lower bound value a.

- There is a lower bound on the possible values for the variance of an unbiased estimator due to Cramér and Rao:

- Theorem (Cramér-Rao Inequality): Suppose that $p_X(x;\theta)$ is a probability density function that is differentiable in θ. Also suppose that the support of p, the set of values of x where $p_X(x;\theta) \neq 0$, does not depend on the parameter θ. If $x_1, x_2, \ldots, x_n$ is a random sample drawn from X, $\hat{\theta} = f(x_1, \ldots, f_n)$ is an unbiased estimator of θ, and $\ell = \ln[p_X(x;\theta)]$ denotes the log-pdf of the distribution, then $\text{var}(\hat{\theta}) \geq 1/I(\theta)$ where $I(\theta) = n \cdot E[(\partial\ell/\partial\theta)^2]$.

  - In the event that p_X is twice-differentiable in θ, it can be shown that $I(\theta)$ can also be calculated as $I(\theta) = -n \cdot E[\partial^2\ell/\partial\theta^2]$.
  - The proof of this theorem is rather technical (although not conceptually difficult), so we will omit the precise details, other than to remark that the key detail yielding the actual inequality is based on the fact that for two random variables S and T, we have $\text{var}(S)\text{var}(T) \geq [\text{cov}(S,T)]^2$.
  - In practice, it is not always so easy to evaluate the lower bound in the Cramér-Rao inequality.
  - Furthermore, there does not always exist an unbiased estimator that actually achieves the Cramér-Rao bound. However, if we are able to find an unbiased estimator whose variance does achieve the Cramér-Rao bound, then the inequality guarantees that this estimator is the most efficient possible.

- Example: Suppose that a coin with unknown probability θ of landing heads is flipped n times, yielding results $x_1, x_2, \ldots, x_n$ (where we interpret heads as 1 and tails as 0). Show that the estimator $\hat{\theta} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ is unbiased and has the minimum variance of all possible unbiased estimators.

  - We have $E(x_i) = \theta$ and so $E(\hat{\theta}) = \dfrac{1}{n}[E(x_1) + \cdots + E(x_n)] = \dfrac{1}{n} \cdot n \cdot \theta = \theta$, so $\hat{\theta}$ is unbiased.
  - For the variance, we first compute the Cramér-Rao bound: the likelihood function can be written as $L(x;\theta) = \theta^x(1-\theta)^{1-x}$ (it is θ if $x = 1$ and $1 - \theta$ if $x = 0$), so that $\ell = x\ln\theta + (1-x)\ln(1-\theta)$.
  - Differentiating twice yields $\dfrac{\partial^2\ell}{\partial\theta^2} = -\dfrac{x}{\theta^2} + \dfrac{1-x}{(1-\theta)^2}$, and so since $E(x) = \theta$, the expected value is $E[\dfrac{\partial^2\ell}{\partial\theta^2}] = \dfrac{E(x)}{\theta^2} + \dfrac{E(1-x)}{(1-\theta)^2} = -\dfrac{\theta}{\theta^2} + \dfrac{1-\theta}{(1-\theta)^2} = -\dfrac{1}{\theta(1-\theta)}$.
  - Thus, $I(\theta) = \dfrac{n}{\theta(1-\theta)}$ so the Cramér-Rao bound gives $\text{var}(\hat{\theta}) \geq \dfrac{\theta(1-\theta)}{n}$.
  - But now, $x_1 + x_2 + \cdots + x_n$ is binomially distributed so its variance is $n\theta(1-\theta)$: then the variance of the given estimator $\hat{\theta} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ is $\dfrac{1}{n^2} \cdot n\theta(1-\theta) = \dfrac{\theta(1-\theta)}{n}$, which is precisely the Cramér-Rao bound.
  - This means that our unbiased estimator $\hat{\theta} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ has the minimum possible variance, as claimed.
  - Remark: This result tells us that the best possible estimator for the probability that the coin lands heads is in fact the obvious one, namely, the total proportion of the flips that actually landed heads.

- Example: Show that the maximum-likelihood estimator $\hat{\theta}_\mu = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ is the most efficient possible unbiased estimator of the mean of a normal distribution with unknown mean $\theta = \mu$ and known standard deviation σ.

  - We will show that this estimator achieves the Cramér-Rao bound.
  - For this, we first compute the log-pdf $\ell = -\ln(\sqrt{2\pi}) - \frac{1}{2}\ln(\sigma) - \dfrac{(x-\theta)^2}{2\sigma^2}$.

- Differentiating yields $\frac{\partial \ell}{\partial \theta} = -\frac{x-\theta}{\sigma^2}$ and then $\frac{\partial^2 \ell}{\partial \theta^2} = \frac{1}{\sigma^2}$. Since this is constant we simply see $E[\frac{\partial^2 \ell}{\partial \theta^2}] = \frac{1}{\sigma^2}$, and so the Cramér-Rao bound dictates that $\text{var}(\hat{\theta}) \geq \sigma^2/n$ for any estimator $\hat{\theta}$.
- For our estimator, since the x_i are all independent and normally distributed with mean θ and standard deviation σ, we have $\text{var}(\hat{\theta}_\mu) = \frac{1}{n^2}[\text{var}(x_1) + \cdots + \text{var}(x_n)] = \frac{1}{n^2} \cdot n\sigma^2 = \sigma^2/n$.
- Thus, the variance of our estimator $\hat{\theta}_\mu$ achieves the Cramér-Rao bound, meaning that it is the most efficient unbiased estimator possible.

- Example: Show that the maximum-likelihood estimator $\hat{\theta}_{\text{var}} = \frac{1}{n}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2\right]$ is the most efficient possible estimator of the variance $\theta = \sigma^2$ of a normal distribution with known mean μ and unknown standard deviation σ.
  - We will show that this estimator achieves the Cramér-Rao bound.
  - For this, we first compute the log-pdf $\ell = -\ln(\sqrt{2\pi}) - \frac{1}{2}\ln(\theta) - \frac{(x-\mu)^2}{2\theta}$.
  - Differentiating yields $\frac{\partial \ell}{\partial \theta} = -\frac{1}{2\theta} + \frac{(x-\mu)^2}{2\theta^2}$ and then $\frac{\partial^2 \ell}{\partial \theta^2} = \frac{1}{2\theta^2} - \frac{(x-\mu)^2}{\theta^3}$.
  - Therefore, $E[\frac{\partial^2 \ell}{\partial \theta^2}] = \frac{1}{2\theta^2} - \frac{E[(x-\mu)^2]}{\theta^3}$, and since $E[(x-\mu)^2] = \text{var}(x) = \sigma^2 = \theta$, we obtain $E[\frac{\partial^2 \ell}{\partial \theta^2}] = \frac{1}{2\theta^2} - \frac{\theta}{\theta^3} = -\frac{1}{2\theta^2}$.
  - Therefore, $I(\theta) = -nE[\partial^2 \ell/\partial\theta^2] = n/(2\theta^2)$, and so the Cramér-Rao bound dictates that $\text{var}(\hat{\theta}) \geq 2\theta^2/n$ for any estimator $\hat{\theta}$.
  - For our estimator, first we note that $\text{var}[(x_i-\mu)^2] = E[(x_i-\mu)^4] - E[(x_i-\mu)^2]^2 = 3\sigma^4 - (\sigma^2)^2 = 2\theta^2$, by some direct calculations with the normal distribution that we will omit.
  - Then, since the random variables $(x_i-\mu)^2$ are independent, by properties of variance we have $\text{var}(\hat{\theta}_{\text{var}}) = \frac{1}{n^2}\text{var}[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2] = \frac{1}{n^2} \cdot [\text{var}[(x_1-\mu)^2] + \cdots + \text{var}[(x_n-\mu)^2]] = \frac{1}{n^2} \cdot n \cdot 2\theta^2 = \frac{2\theta^2}{n}$.
  - Thus, the variance of our estimator $\hat{\theta}_{\text{var}}$ achieves the Cramér-Rao bound, meaning that it is the most efficient unbiased estimator possible.

- In certain situations, the Cramér-Rao bound does not apply, and thus we can find estimators that have a smaller variance than it would predict:

- Example: Find the variance of the unbiased estimator $\hat{\theta}_1 = \frac{n+1}{n}\max(x_1, x_2, \ldots, x_n)$ from sampling the uniform distribution on $[0, \theta]$, and compare it to the Cramér-Rao bound.
  - We already computed $\text{var}(\hat{\theta}_1) = \frac{1}{n(n+2)}\theta^2$ earlier.
  - To compute the bound from Cramér-Rao, we have $L(\theta) = (1/\theta)^n$ hence $\ell = \ln(L) = -n\ln(\theta)$. Then $\partial\ell/\partial\theta = -\frac{n}{\theta}$ so $\partial^2\ell/\partial\theta^2 = \frac{n}{\theta^2}$. Since this is constant, we simply get $I(\theta) = -n \cdot E[\partial^2\ell/\partial\theta^2] = \frac{n^2}{\theta^2}$, and so the Cramér-Rao bound is $\text{var}(\hat{\theta}) \geq \frac{1}{n^2}\theta^2$.
  - But now notice that $\text{var}(\hat{\theta}_1) < \frac{1}{n^2}\theta^2$: this means $\hat{\theta}_1$ actually has a *smaller* variance than the Cramér-Rao minimum!
  - This is not a contradiction, because in fact one of the hypotheses of the Cramér-Rao theorem is violated: specifically, the condition that the set of values of x where $p_X(x;\theta) \neq 0$ does not depend on the parameter θ. Here, $p_X(x;\theta) \neq 0$ for $x \in [0, \theta]$, and this range clearly does depend on θ.

3.2 Interval Estimation

- When estimating an unknown parameter, it is (of course) desirable to have a prediction that is as accurate as possible. However, it is also important to be able to describe how accurate the prediction is, which is to say, how much the prediction differs from the true parameter value.
    - For example, if we are estimating the height of a building, an estimate of 25.43 meters is certainly useful, but it is far more useful if we can also say that it is correct to within 0.01 meters. (In contrast, if we estimate the height to be 25.43 meters but only to within 20 meters, the estimate is not nearly as good!)
    - If we measure the height three times and obtain estimates of 25.43 meters, 25.41 meters, and 25.44 meters, we can be more confident in the overall accuracy than if the three measurements were 25.43 meters, 17.42 meters, and 33.15 meters. Nonetheless, these measurements by themselves do not provide an explicit error range for our estimated height.

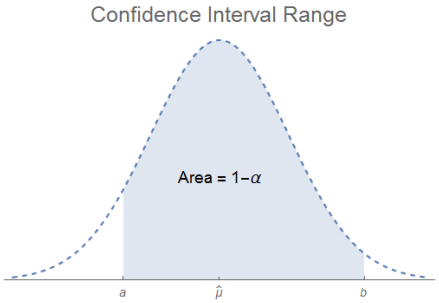
3.2.1 Confidence Intervals

- What we are seeking is to expand our discussion from pointwise parameter estimates, where we estimate the actual value of the parameter, to interval estimates, where we give an interval that we believe the parameter should lie in.
    - Our discussion of unbiasedness is partially in this direction, since unbiasedness eliminates the existence of systematic error (i.e., error that tends to bias the estimate either too high or too low on average).
    - Our discussion of estimator efficiency also represents partial progress toward this goal: efficient estimators have a smaller variance, and thus (by definition) will display less variation in their values than less-efficient estimators.
    - However, unbiasedness is only an average measure, while efficiency is a measurement of precision (the closeness of the measurements to one another) rather than of accuracy (the closeness of the measurements to the true value).
    - What we are seeking is a way to quantify the accuracy of our estimations.
- One approach to quantifying the uncertainty in our estimates is to construct a confidence interval: this is an interval around our estimated value in which we believe the true value should lie.
    - Of course, since the estimator is itself defined in terms of the values of a random sample, we cannot generally be completely certain that the true value of the parameter lies in any useful interval we could define. (We could, of course, simply declare our confidence interval to be the entire real line, but this would not give a useful prediction!)
    - But what we can do is compute the probability that the true parameter value lies in the interval we give. If the probability is reasonably large (depending on the context, one may consider values such as 50%, or 90%, or 95%, or 99% as appropriately large probabilities) then we can be reasonably confident in the accuracy of our estimation.
- Definition: If X is a random variable and $0 < \alpha < 1$, a $100(1-\alpha)\%$ confidence interval for X is an interval (a, b) with $a < X < b$ such that $P(a < X < b) = 1 - \alpha$.
    - We use the notation $100(1-\alpha)\%$ is because it is traditional to quote the size of the confidence interval as a percent, rather than as a raw probability. Thus, for example, a 95% confidence interval for X is an interval (a, b) where X should land 95% of the time.
    - In principle, one can define confidence intervals for any random variable, but in practice they are only given for random variables that represent estimators of unknown parameters.
    - When θ is an unknown parameter, we interpret a confidence interval for θ as giving us a "reasonable error range" (for a precisely quantified notion of reasonable, determined by the error probability α) on a specific estimation $\hat{\theta}$ for θ that we have computed.

- Example: Suppose we perform a maximum likelihood estimate for the parameter $\lambda = \theta$ of a Poisson distribution and obtain the estimate $\hat{\theta} = 1.39$, and by analysis of the variation of the estimator we are able to determine that there is a 95% probability that the true value of θ lies in the interval $(1.33, 1.51)$.
  - This interval $(1.33, 1.51)$ is then a 95% confidence interval for our estimate, and it provides substantial additional context to our pointwise estimate $\hat{\theta} = 1.39$, since it quantifies how much variation we should expect to see in the true value of the parameter.
  - If we sampled this distribution repeatedly and constructed a 95% confidence interval using each sample, we would expect the true value of the parameter to lie inside the interval 95% of the time.
- When we are constructing confidence intervals using parameter estimates, we typically will want to work with unbiased estimators that are as efficient as possible.
  - If the estimator is unbiased, then the confidence interval will not tend to be biased above or below the true value of the parameter (i.e., it yields better average accuracy from a given data sample).
  - If the estimator is efficient, then the size of the interval will be as small as possible, which yields tighter estimates for a given confidence level (i.e., it yields better overall precision from a given data sample).
- In general, computing a confidence interval requires being able to analyze the precise nature of the variation in the estimator $\hat{\theta}$ relative to the true value θ.
  - In certain situations, we can describe this variation quite precisely, but in others it can be very difficult.

3.2.2 Normal Confidence Intervals

- To illustrate one of the simplest cases of computing a confidence interval, suppose we sample a normal distribution with unknown mean μ and known standard deviation σ, obtaining values $x_1, x_2, \ldots, x_n$: our goal is to give a confidence interval for μ.
  - We have previously shown that the maximum likelihood estimator for the mean, which is simply the sample mean $\hat{\mu} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$, is unbiased and is the most efficient unbiased estimator for μ.
  - Furthermore, from our results on the normal distribution and the central limit theorem, we know that since the x_i are independent and normally distributed with mean μ and standard deviation σ, the sample mean $\hat{\mu} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ will also be normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.
  - So far, we have proceeded as if we knew μ and wanted to understand the variation in $\hat{\mu}$.
  - But now we can switch our focus from the variation of $\hat{\mu}$ given μ to the variation of μ given $\hat{\mu}$: from above, the difference $\hat{\mu} - \mu$ is normally distributed with mean 0 and standard deviation $\sigma/\sqrt{n}$.
  - This is the same as saying that the value of μ is normally distributed with mean $\hat{\mu}$ and standard deviation $\sigma/\sqrt{n}$.
  - From this last statement, we can easily derive confidence intervals for the unknown parameter μ using properties of the normal distribution.
  - Explicitly, if $N_{\hat{\mu},\sigma/\sqrt{n}}$ is the normal distribution with mean $\hat{\mu}$ and standard deviation $\sigma/\sqrt{n}$, then $P(a < \mu < b) = P(a < N < b)$.
  - We can therefore construct a $100(1-\alpha)\%$ confidence interval for μ simply by finding a range (a, b) such that $P(a < N_{\hat{\mu},\sigma/\sqrt{n}} < b) = 1 - \alpha$, as illustrated in the diagram below:



- There are many possible choices for this interval, so to narrow things down, we usually require that the interval be symmetric around $\hat{\mu}$, and for convenience we often also rephrase this condition in terms of the standard normal distribution $N_{0,1}$ by rescaling.
- If we compute the constant $z_{\alpha/2}$ such that $P(-z_{\alpha/2} < N_{0,1} < z_{\alpha/2}) = 1-\alpha$, then this yields the $100(1-\alpha)\%$ confidence interval $(a,b) = (\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.
- Using the symmetry of the normal distribution, $P(-z_{\alpha/2} < N_{0,1} < z_{\alpha/2}) = 1-\alpha$ is equivalent to $P(N_{0,1} < -z_{\alpha/2}) = \alpha/2$, or also to $P(z_{\alpha/2} < N_{0,1}) = 1-(\alpha/2)$, which allows us to compute the value of $z_{\alpha/2}$ by evaluating the inverse cumulative distribution function for $N_{0,1}$.
- Indeed, this is why we used the notation $z_{\alpha/2}$, since it is essentially just the value of the inverse cumulative distribution function for $N_{0,1}$ evaluated at $\alpha/2$, up to a minus sign.

- We can summarize the results of this discussion as follows:
- Proposition (Normal Confidence Intervals): A $100(1-\alpha)\%$ confidence interval for the unknown mean μ of a normal distribution with known standard deviation σ is given by $\hat{\mu} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ where n sample points $x_1, \ldots, x_n$ are taken from the distribution, $\hat{\mu} = \frac{1}{n}(x_1 + \cdots + x_n)$ is the sample mean, and c is the constant satisfying $P(-z_{\alpha/2} < N_{0,1} < z_{\alpha/2}) = 1-\alpha$.
  - Some specific values of $z_{\alpha/2}$ for various common values of α are given in the table below:

$1-\alpha$	50%	80%	90%	95%	98%	99%	99.5%	99.9%
$z_{\alpha/2} : P(-z_{\alpha/2} < N_{0,1} < z_{\alpha/2}) = 1-\alpha$	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.2905

  - The second term $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is often called the margin of error for the confidence interval, since it represents the maximum distance away (in either direction) values in the interval can be from the center of the interval.
  - If we imagine choosing different sample sizes n, we can see that the margin of error in the estimate decreases with larger n. This is, of course, quite intuitive: if we sample more values, we would expect the errors to tend to cancel one another out on average, yielding an average that is more likely to land close to the true value than any single observation. (More formally, it follows from the central limit theorem.)
  - More precisely, the margin of error will be proportional to $1/\sqrt{n}$: so, for example, to cut the margin of error in half would require a sample size that is 4 times as large.

- Example: A normal distribution with unknown mean μ and standard deviation $\sigma = 1$ is sampled four times, yielding the values 1.4, 0.2, 2.9, and 1.1. Find 50%, 90%, 95%, and 99% confidence intervals for μ.
  - Here, we have $n = 4$, $\hat{\mu} = \frac{1}{4}(1.4 + 0.2 + 2.9 + 1.1) = 1.4$, and $\sigma/\sqrt{n} = 0.5$.
  - From the proposition and the table of values below it, we obtain the 50% confidence interval $\hat{\mu} \pm 0.6745 \cdot \sigma/\sqrt{n} = \boxed{(1.063, 1.737)}$, the 90% confidence interval $\hat{\mu} \pm 1.6449 \cdot \sigma/\sqrt{n} = \boxed{(0.577, 2.223)}$, the 95% confidence interval $\hat{\mu} \pm 1.9600 \cdot \sigma/\sqrt{n} = \boxed{(0.420, 2.380)}$, and the 99% confidence interval $\hat{\mu} \pm 2.5758 \cdot \sigma/\sqrt{n} = \boxed{(0.112, 2.688)}$.

- Example: From analysis of industrial fabrication, it is determined that the diameters of bolts manufactured at Factory X are distributed normally with mean 20mm and standard deviation 0.01mm. To check the quality of each manufacturing lot, a random sample of bolts are selected and their average diameter is measured. Assuming the standard deviation is 0.01mm, if 10 bolts are selected and the average diameter is 20.0144mm, find 50%, 90%, 95%, and 99% confidence intervals for the true mean of the lot. Based on these calculations, is it likely that the true mean is actually 20mm?
  - Here, we have $n = 10$, $\hat{\mu} = 20.0144\text{mm}$, and $\sigma/\sqrt{n} = 0.01/\sqrt{10} = 0.00316\text{mm}$.
  - From the proposition and the table of values below it, we obtain the 50% confidence interval $\hat{\mu} \pm 0.6745 \cdot \sigma/\sqrt{n} = \boxed{(20.0123\text{mm},\ 20.0165\text{mm})}$, the 90% confidence interval $\hat{\mu} \pm 1.6449 \cdot \sigma/\sqrt{n} = \boxed{(20.0092\text{mm}, 20.0196\text{mm})}$, the 95% confidence interval $\hat{\mu} \pm 1.9600 \cdot \sigma/\sqrt{n} = \boxed{(20.0082\text{mm}, 20.0206\text{mm})}$, and the 99% confidence interval $\hat{\mu} \pm 2.5758 \cdot \sigma/\sqrt{n} = \boxed{(20.0063\text{mm}, 20.0225\text{mm})}$.

- Based on these confidence intervals, it seems very unlikely that the true mean is actually 20mm: even a 99% confidence interval does not contain this value.
- Despite the fact that the average diameter of the bolts in the sample only differs from the desired one by 0.0144mm (which is 1.44 times the standard deviation of the bolt diameter), this is in fact very strong evidence that the true mean of this lot of bolts is not actually 20mm. In the next chapter, we will extend this type of analysis to describe methods for testing the hypothesis that the bolt diameter is actually equal to 20mm.

- Example: In the same scenario as above, if instead a 99% margin of error of at most 0.005mm for the true mean diameter is desired, what is the minimum number of bolts that should be sampled to achieve this level of precision?
  - The margin of error for a 99% confidence interval in this setting is $2.5758 \cdot \sigma/\sqrt{n}$. Since this quantity is required to be 0.005mm, solving the resulting equation for n gives $n = \left(\dfrac{2.5758 \cdot 0.01\text{mm}}{0.005\text{mm}}\right)^2 \approx 26.54$.
  - This means a sample of $\boxed{27}$ bolts would be sufficient to give the desired precision.

- Example: A marine biologist measures the lengths of 100 adult blue whales and, in his sample, the average length was 27.11m with a standard deviation of 1.3m. Assuming that this standard deviation is correct for the full population, find (i) a 98% confidence interval for the average length of a blue whale, (ii) the number of blue whales that would need to be measured to give a 98% confidence interval with half the margin of error, and (iii) the probability that if another 100 blue whales were independently sampled, both 98% confidence intervals would contain the true mean.
  - For (i), we have $n = 100$, $\hat{\mu} = 27.11\text{m}$, and $\sigma/\sqrt{n} = 1.3\text{m}/\sqrt{100} = 0.13\text{m}$.
  - Thus, using the table, we obtain the 98% confidence interval $(\hat{\mu} - 2.3263 \cdot \sigma/\sqrt{n},\ \hat{\mu} + 2.3263 \cdot \sigma/\sqrt{n}) = \boxed{(26.81\text{m},\ 27.41\text{m})}$.
  - For (ii), since the width of the confidence interval is $2.3263 \cdot \sigma/\sqrt{n}$, to halve the width we would require a value of n four times as large, which is $n = \boxed{400}$.
  - For (iii), by definition each confidence interval has a probability 0.98 of containing the true mean. Since these samples are independent, the probability that both contain the true mean is $(0.98)^2 = \boxed{0.9604}$.

- We will remark that the assumption in the previous example, that the sample standard deviation is equal to the population standard deviation, is not generally valid in practice, and leads to narrower confidence intervals than those derived by taking the difference between the sample standard deviation and the population standard deviation into account.
  - Above, our discussion effectively analyzes the ratio $\dfrac{x - \overline{x}}{\sigma/\sqrt{n}}$ where σ is the (known) population standard deviation by observing that this random variable has a standard normal distribution.
  - However, if we replace the population standard deviation σ by the sample standard deviation S, the resulting ratio $\dfrac{x - \overline{x}}{S/\sqrt{n}}$ is no longer normally distributed. (We can therefore not construct confidence intervals using the procedure described above.)
  - As we discuss in a later chapter, the random variable $\dfrac{x - \overline{x}}{S/\sqrt{n}}$ actually follows a distribution known as the t distribution.

- Example: The weight of domestic housecats is normally distributed with a population average of 4.02kg and a standard deviation of 0.24kg. Some cats from two feral colonies, Colony A and Colony B, are each weighed. The 16 cats from Colony A had an average weight of 3.95kg with a standard deviation of 0.30kg, while the 9 cats from Colony B had an average weight of 4.24kg with a standard deviation of 0.37kg. Find 90% confidence intervals for (i) the average weight of cats in each colony, (ii) the total weight of another equally-sized sample of cats from each colony, and (iii) for the difference between the average weights in the two colonies.

- For Colony A, we have $n = 16$, $\hat{\mu} = 3.95$kg, and $\sigma/\sqrt{n} = 0.06$kg. (Note that the standard deviation of the sample is irrelevant, because we are given the standard deviation $\sigma = 0.24$kg of the population.)
- Thus, we have a 90% confidence interval for the average weight given by $\hat{\mu} \pm 1.6449\sigma/\sqrt{n} = \boxed{(3.85\text{kg}, 4.05\text{kg})}$.
- For the total weight, if we resample with a different 16 cats, we simply scale the interval by 16, yielding the 90% confidence interval $\boxed{(61.6\text{kg}, 64.8\text{kg})}$.
- In the same way, for colony B, we have $n = 9$, $\hat{\mu} = 4.24$kg, and $\sigma/\sqrt{n} = 0.08$kg.
- Thus, we have a 90% confidence interval for the average weight given by $\hat{p} \pm 1.6449\sigma/\sqrt{n} = \boxed{(4.11\text{kg}, 4.37\text{kg})}$.
- For the total weight, if we resample with 9 cats, we simply scale the interval by 9, yielding the 90% confidence interval $\boxed{(37.0\text{kg}, 39.3\text{kg})}$.
- The difference in the average weights is a little bit trickier. The idea is to observe that if A is normally distributed with mean μ_A and standard deviation σ_A and B is normally distributed with mean μ_B and standard deviation σ_B, then $B - A$ is normally distributed with mean $\mu_B - \mu_A$ and standard deviation $\sqrt{\sigma_A^2 + \sigma_B^2}$ (the latter is because the variance is additive).
- Thus, from the analysis above, the difference in the average weights is normally distributed with standard deviation $\sqrt{(0.06\text{kg})^2 + (0.08\text{kg})^2} = 0.1$kg. Since the observed average is $4.24\text{kg} - 3.95\text{kg} = 0.29$kg, by our discussion the 90% confidence interval for the difference in the average weights will be $0.29\text{kg} \pm 1.6446 \cdot 0.1\text{kg} = \boxed{(0.125\text{kg}, 0.454\text{kg})}$.

3.2.3 Binomial Confidence Intervals

- For distributions that are well approximated by the normal distribution, we can use methods similar to those for the normal distribution to obtain very accurate approximate confidence intervals.
- One particularly important situation of interest is the case of the binomial distribution, which arises from repeated sampling of a Bernoulli random variable.
  - So, suppose that we have a Bernoulli random variable with success probability p that we sample n times, yielding sample values $x_1, x_2, \ldots, x_n$ with a total number of successes equal to $k = x_1 + x_2 + \cdots + x_n$.
  - Then as we have shown, the sample success estimator $\hat{p} = k/n$ is unbiased and is the most efficient possible unbiased estimator of the true success probability p.
  - Furthermore, the sample estimator $n\hat{p}$ (which counts the total number of successes in the n samples) will be binomially distributed with mean np and standard deviation $\sqrt{np(1-p)}$.
  - To compute an exact confidence interval, we would need to determine the precise nature of the relationship between $\hat{p} = k/n$ and the parameter p itself, which is quite difficult to do directly.
  - However, when np and $n(1-p)$ are both reasonably large, the binomial distribution will be well approximated by the corresponding normal distribution, and so $n\hat{p}$ will have an approximately normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$.
  - Equivalently, this says $\hat{p}$ will have an approximately normal distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.
  - We can now invert our focus and switch from using p to study the variation in $\hat{p}$ to using $\hat{p}$ to study the variation in p.
  - However, there is one crucial difference: in our earlier setup, we were given the standard deviation of the distribution explicitly. Here, the standard deviation still depends on the (now) unknown parameter p.
  - To avoid having to study the complicated way in which the exact distribution of p would depend on $\hat{p}$, we also make the simplifying assumption[4] that $\hat{p}$ is a sufficiently good estimate of p that $\sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n}$.

[4] This is a fairly reasonable assumption because (as we just observed) $\hat{p}$ is roughly normally distributed with mean p and standard deviation $\sqrt{p(1-p)/n}$, which is typically much smaller than p. Indeed, 99.8% of the time, $|\hat{p} - p| < 3\sqrt{p(1-p)/n}$, and then one may linearize the difference in the two standard deviations to get the estimate $\left|\sqrt{\hat{p}(1-\hat{p})/n} - \sqrt{p(1-p)/n}\right| \approx |3 - 6\hat{p}|/(2n) + O(n^{-2})$. This is quite small relative to the actual standard deviation, especially when n is large, which will always be the case when we are using the normal approximation to the binomial distribution.

- Thus, our approximation says that the distribution of p is approximately normal with mean $\hat{p}$ and standard deviation $\sqrt{\hat{p}(1-\hat{p})/n}$. Now, at last, we can apply our analysis of the normal distribution to construct confidence intervals for p:

- Proposition (Binomial Confidence Intervals): Suppose a Bernoulli random variable is sampled n times yielding k successes, for an overall sample success rate of $\hat{p} = k/n$. In situations where the normal approximation to the binomial distribution is accurate (heuristically, when k and $n-k$ are both larger than 5 or so), then a $100(1-\alpha)\%$ confidence interval for the true success probability p is given by $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = (\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p}+z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n})$, where $z_{\alpha/2}$ is the constant satisfying $P(-z_{\alpha/2} < N_{0,1} < z_{\alpha/2}) = 1-\alpha$.
  - More compactly, this result says that our best estimate for the overall success rate is $\hat{p} = k/n$, and the margin of error at the $100(1-\alpha)\%$ confidence level is equal to $z_{\alpha/2}\sigma$ where $\sigma = \sqrt{\hat{p}(1-\hat{p})/n}$ is the sample proportion's standard deviation.
  - We repeat our table of specific values of $z_{\alpha/2}$ for various common values of α:

$1-\alpha$	50%	80%	90%	95%	98%	99%	99.5%	99.9%
$z_{\alpha/2} : P(-z_{\alpha/2} < N_{0,1} < z_{\alpha/2}) = 1-\alpha$	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.2905

- Example: A coin with unknown probability p of landing heads is flipped 100 times, yielding 64 heads. Find 50%, 80%, 90%, and 99.5% confidence intervals for p. How likely does it seem, based on these calculations, that the coin is actually fair?
  - Here, we have $n = 100$ and $\hat{p} = 64/100 = 0.64$, so that $\sigma = \sqrt{\hat{p}(1-\hat{p})/n} = 0.048$.
  - Thus, from the proposition and the table of values, we obtain the 50% confidence interval $\hat{p} \pm 0.6745\sigma = \boxed{(0.6076, 0.6724)}$, the 80% confidence interval $\hat{p} \pm 1.2816\sigma = \boxed{(0.5785, 0.7015)}$, the 90% confidence interval $\hat{p} \pm 1.6449\sigma = \boxed{(0.5610, 0.7190)}$, and the 99.5% confidence interval $\hat{p} \pm 2.0870\sigma = \boxed{(0.5053, 0.7747)}$.
  - Even with the 99.5% confidence interval, the value 0.5 is not contained inside the confidence interval (though it is not that far below the lower bound): this suggests it is very unlikely that the coin was actually fair.
  - As a sanity check, the probability of obtaining 64 heads when actually flipping a fair coin is $\binom{100}{64}/2^{100} \approx 0.156\%$, whereas if the heads probability was actually 0.64 then the probability of obtaining 64 heads would be $\binom{100}{64}(0.64)^{64}(0.36)^{36} \approx 8.288\%$: quite a lot more likely. This is precisely what our maximum likelihood estimate for $\hat{p}$ confirms: the value $p = 0.64$ is the one where were are the most likely to observe 64 heads in 100 flips.

- Example: Last season, a professional basketball player attempted 1800 two-point shots and made 753 of them. Find (i) 80% and 99% confidence intervals for his shooting average last season, and (ii) 80% and 99% confidence intervals for the total number of shots he should expect to make this season if he attempts 1500 shots and his true shooting average stays the same as last season.
  - For (i), we have $n = 1800$ and $\hat{p} = 753/1800 \approx 41.83\%$, so that $\sigma = \sqrt{\hat{p}(1-\hat{p})/n} \approx 1.163\%$.
  - Using the table, we obtain the 80% confidence interval $\hat{p} \pm 1.2816\sigma = \boxed{(40.34\%,\ 43.32\%)}$ and the 99% confidence interval $\hat{p} \pm 2.5758\sigma = \boxed{(38.83\%,\ 44.83\%)}$.
  - For (ii), the distribution of the total number of made shots out of 1500 attempts will be binomial with mean $1500p$ and standard deviation $\sqrt{1500p(1-p)}$.
  - Using the approximation $\sqrt{p(1-p)} \approx \sqrt{\hat{p}(1-\hat{p})}$ (which we also used, and justified as reasonable, in our analysis of the binomial distribution above) and approximating the binomial distribution with the normal distribution of the same mean and standard deviation, we would expect that the number of made shots this season is distributed approximately normally with mean $1500p$ and standard deviation $\sigma' = \sqrt{1500\hat{p}(1-\hat{p})} = 19.105$.
  - Therefore, taking the parameter estimate $1500\hat{p} = 627.5$ yields the 80% confidence interval $1500\hat{p} \pm 1.2816\sigma' = \boxed{(603,\ 652)}$ and the 99% confidence interval $1500\hat{p} \pm 2.5758\sigma' = \boxed{(578, 677)}$.

- An extremely common use of confidence intervals is in polling statistics, where a random sample of a population is used to estimate the proportion that support a particular measure.
  - Typically, most polls report the margin of error associated with a 95% confidence interval. In popular parlance, it is usually referred to as simply "the margin of error", with no qualifier, but most reputable polls also include the confidence level with their statistics.
  - Thus, if a poll reports "45% of voters support X, with a margin of error of 6%" then this typically means that the 95% confidence interval for the percent support of X is $(39\%, 51\%)$.
  - It is important not to misinterpret the meaning of the confidence interval above: although a portion of the confidence interval does include outcomes where the support of X is above 50%, it is far more likely that the support for X is below 50% than above 50%.
- Example: A pollster wishes to measure the statewide support for Proposition Q. He randomly samples 1000 likely voters and finds 540 of them support Proposition Q. Find 95% and 99.9% confidence intervals, and the associated margins of error, for the true percentage of the population that supports Proposition Q.
  - Here, we have $n = 1000$ and $\hat{p} = 540/1000 = 54\%$, so that $\sigma = \sqrt{\hat{p}(1-\hat{p})/n} \approx 1.576\%$.
  - Thus, from the proposition and the table of values below it, we obtain the 95% confidence interval $\hat{p} \pm 1.9600\sigma = \boxed{(50.91\%,\ 57.09\%)}$, and the 99.9% confidence interval $\hat{p} \pm 3.2905\sigma = \boxed{(48.81\%,\ 59.19\%)}$.
  - The margin of error for the 95% confidence interval is $1.9600\sigma \approx \boxed{3.09\%}$, and the margin of error for the 99.9% confidence interval is $3.2905\sigma \approx \boxed{5.19\%}$.
- Example: A pollster wishes to measure the support for the statewide support of Proposition R. If she expects the support level for the proposition to be approximately 65%, what is the smallest number of people needed for the 95% confidence interval's margin of error to be at most $\pm 2\%$? How would the answer change if the support level for the proposition is unknown?
  - Here, the expected proportion is $\hat{p} = 0.65$, so $\sigma = \sqrt{\hat{p}(1-\hat{p})/n}$. At the 95% confidence level, the margin of error is 1.9600σ; setting this equal to 2% and solving yields $n = \dfrac{\hat{p}(1-\hat{p})}{(0.02/1.9600)^2} \approx 2184.9$.
  - Thus, the minimum number of people needed for the poll will be $\boxed{2185}$ to achieve a 2% margin of error at the 95% confidence level.
  - If the support level $\hat{p}$ is unknown, the largest possible value of n will occur when the numerator $\hat{p}(1-\hat{p})$ is maximized, which (either by calculus or completing the square) occurs for $\hat{p} = 1/2$.
  - The corresponding value of n is then $\dfrac{(1/2)\cdot(1/2)}{(0.02/1.9600)^2} \approx 2401.02$, which rounds up to $\boxed{2402}$ people.
- Example: A political article states "Based on a recent poll, candidate Y has an approval rating of $43.1\% \pm 3.6\%$ (95% CI, $n = 750$), so it is impossible for their true approval to be 50% or above". Critique this statement.
  - Because the poll was conducted by sampling, there is always a possibility (however remote) that the actual favorability rating lies outside any given confidence interval.
  - In this case, given that the sample size was $n = 750$ and that it is a 95% confidence interval with a success probability $\hat{p} = 0.431$, the actual distribution of the true favorability rating will be normal with mean 42.1% and standard deviation $\sqrt{\hat{p}(1-\hat{p})/n} \approx 1.81\%$. (Note that this is consistent with the quoted information since the margin of error would then be $1.9600\sigma \approx 3.54\%$.)
  - Using properties of the normal distribution, we can then compute $P(N_{43.1,1.81} > 50) = P(N_{0,1} > 3.816) \approx 0.068\%$. So, although it is fairly unlikely that candidate Y's favorability rating is actually 50% or above, it is certainly still possible. Note that we are assuming here that the poll was an independent and truly random sample: if the selection of respondents was nonrandom (e.g., if they were biased against the candidate), the results could deviate substantially from reality.

Well, you're at the end of my handout. Hope it was helpful.