19 Applications of Tail Bounds: Confidence Intervals and Balls and Bins

In Chapter 18 we saw several powerful tail bounds, including the Chebyshev bound and the Chernoff bound. These are particularly useful when bounding the tail of a sum of independent random variables. We also reviewed the application of the Central Limit Theorem (CLT) to approximating the tail of a sum of independent and identically distributed (i.i.d.) random variables.

These tail bounds and approximations have immediate application to the problem of interval estimation, also known as creating "confidence intervals" around an estimation. They also are very useful in solving an important class of problems in theoretical computer science, called "balls and bins" problems, where balls are thrown at random into bins. Balls-and-bins problems are in turn directly related to hashing algorithms and load-balancing algorithms. In this chapter, and the next, we will study these immediate applications of our existing tail bounds and approximations. In Chapters 21–23, we will move on to the topic of randomized algorithms, where we will see many more applications of our tail bounds.

19.1 Interval Estimation

In Chapter 15, we discussed estimating the mean, $\mathbf{E}[X]$, of a random variable (r.v.) X. We assume that we're given n i.i.d. samples of X, which we denote by $X_1, X_2, \ldots, X_n$. We then define our estimator of $\mathbf{E}[X]$ to be

$$\overline{X} \equiv \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We call $\overline{X}$ the **sample mean**. Importantly, $\overline{X}$ is a function of random samples and thus is itself a *random variable*, not a constant.

What we have not discussed, though, is: *How good is $\overline{X}$ at estimating $\mathbf{E}[X]$?*

Clearly, the estimator $\overline{X}$ gets closer and closer to $\mathbf{E}[X]$ as we increase the number of samples n. But it's hard to say how good $\overline{X}$ is because it's just a single value: a point estimator. What we really want is an interval around $\overline{X}$ where we can say that the true mean, $\mathbf{E}[X]$, lies within that interval with high confidence, say 95% probability. That is, we want an "interval estimator."

Definition 19.1 *Let θ be some parameter of r.v. X that we're trying to estimate, e.g., $\mathbf{E}[X]$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of X. Then we say that an* **interval estimator** *of θ with confidence level $1-\alpha$ is a pair of estimators, $\hat{\theta}_{low}$ and $\hat{\theta}_{high}$, where*

$$\mathbf{P}\left\{\hat{\theta}_{low} \leq \theta \leq \hat{\theta}_{high}\right\} \geq 1-\alpha.$$

Importantly, the randomness here is due to $\hat{\theta}_{low}$ and $\hat{\theta}_{high}$, not θ. Here θ is a constant that we're trying to estimate, while $\hat{\theta}_{low}$ and $\hat{\theta}_{high}$ are both functions of the random data samples $X_1, \ldots, X_n$ and hence are random variables. Equivalently, we say that

$$\left[\hat{\theta}_{low}, \hat{\theta}_{high}\right]$$

is a $(1-\alpha)\cdot 100\%$ **confidence interval** *for θ, with* **width** *$\hat{\theta}_{high} - \hat{\theta}_{low}$.*

For the purpose of our discussion we will be looking at creating 95% confidence intervals on $\mathbf{E}[X]$, which will take the form of

$$[\overline{X} - \delta, \overline{X} + \delta],$$

where 2δ represents the width of our confidence interval and $\overline{X}$ is the sample mean. It is generally desirable that the confidence interval has both a *high confidence level* (say 95%) and also a *low width*.

In Section 19.2 we'll see how to develop confidence intervals with guarantees. To do this, we will use Chernoff and Chebyshev bounds. Unfortunately, it is not always possible to develop these "exact" (guaranteed) confidence intervals. In Section 19.3 we show how to develop *approximate* confidence intervals. These rely on the CLT approximation.

19.2 Exact Confidence Intervals

In developing confidence intervals we start with the classical example of polling to determine the outcome of an election. Our goal here is to develop 95% confidence intervals, but this can easily be generalized to any confidence level.

19.2.1 Using Chernoff Bounds to Get Exact Confidence Intervals

Example 19.2 (Polling for election)

Imagine that we are trying to estimate the fraction of people who will vote for

Biden in the presidential election. Let p be the true fraction. Our goal is to figure out p.

To estimate p, we use the following algorithm:

1. Sample $n = 1000$ people independently at random. Let X_i be an indicator r.v., which is 1 if the ith person sampled says they'll vote for Biden.
2. Let $S_n = X_1 + X_2 + \cdots + X_n$.
3. Return the r.v.
$$\overline{X} = \frac{S_n}{n}$$
as our estimate of p.

Question: Why is X_i a r.v.? How is it distributed?

Answer: Each individual either votes for Biden or doesn't, so there's no randomness in a particular individual. The *randomness* comes from the fact that we're picking *random* individuals. If we let X_i be our ith sample, then,
$$X_i = \begin{cases} 1 & \text{if person } i \text{ said yes} \\ 0 & \text{otherwise} \end{cases}.$$

Here $X_i \sim \text{Bernoulli}(p)$, because the probability that a randomly chosen person says "yes" is p.

Question: What do S_n and $\overline{X}$ represent? How are they distributed?

Answer: S_n represents the total number of people sampled who say they'll vote for Biden and $\overline{X}$ represents the fraction of people sampled who say they'll vote for Biden. Both are functions of random variables, so both are random variables.
$$S_n \sim \text{Binomial}(n, p) \qquad \overline{X} \sim \frac{1}{n} \cdot \text{Binomial}(n, p).$$

Our **goal** is to define a 95% confidence interval on p where:
$$\mathbf{P}\left\{p \in [\overline{X} - \delta, \overline{X} + \delta]\right\} \geq 95\%.$$

Question: Given that n people are sampled, and we want a 95% confidence interval on p, how can we frame this as a Chernoff bound problem?

Hint: To use a Chernoff bound, we want to phrase the question as the probability that a Binomial deviates from its mean by some amount.

Answer: We need to find a δ such that

$$\mathbf{P}\left\{\left|\overline{X} - p\right| > \delta\right\} < 5\%, \tag{19.1}$$

or equivalently, such that

$$\mathbf{P}\left\{|S_n - np| > n\delta\right\} < 5\%. \tag{19.2}$$

We're thus considering the probability that S_n deviates from its mean, np, by $n\delta$. By using both parts of the Chernoff bound in Theorem 18.4, we have

$$\mathbf{P}\left\{|S_n - \mathbf{E}\left[S_n\right]| > n\delta\right\} \leq 2e^{-\frac{2(n\delta)^2}{n}}.$$

Hence, we need to find a δ such that

$$2e^{-2n\delta^2} < 0.05,$$

Equivalently,

$$\delta > \sqrt{\frac{-\ln 0.025}{2n}} = \sqrt{\frac{1.84}{n}}. \tag{19.3}$$

Question: How does the width of our confidence interval scale with the number of sampled people?

Answer: Observe that δ scales as $\frac{1}{\sqrt{n}}$. The bigger n is, the smaller δ can be.

Question: If we sample $n = 1000$ people, what is our confidence interval?

Answer: For $n = 1000$, we have $\delta \approx 0.043$. Hence $[\overline{X} - 0.043, \overline{X} + 0.043]$ forms a 95% confidence interval on the true p.

Question: Suppose that we need the width of our confidence interval to be no more than 1%, while still maintaining a 95% confidence level? How can we change n to achieve this?

Answer: We now have *two constraints*:

$$\delta > \sqrt{\frac{1.84}{n}} \qquad \text{and} \qquad \delta \leq 0.005.$$

So,

$$\sqrt{\frac{1.84}{n}} \leq 0.005,$$

or equivalently,

$$n \geq \frac{1.84}{(0.005)^2} = 73{,}600.$$

Of course, there are many more issues that come up in polling estimation. For example, it is not obvious how to get "independent," equally weighted samples.

19.2.2 Using Chebyshev Bounds to Get Exact Confidence Intervals

Question: Let's return to the problem of obtaining a 95% confidence interval on p given n sampled people, but this time we want to use Chebyshev's bound. Can we do it?

Answer: As in (19.2), we again need to find a δ such that

$$\mathbf{P}\{|S_n - np| > n\delta\} < 5\%.$$

By Chebyshev's Inequality (Theorem 18.2),

$$\mathbf{P}\{|S_n - np| > n\delta\} \leq \frac{\mathbf{Var}(S_n)}{(n\delta)^2} = \frac{np(1-p)}{n^2\delta^2}.$$

So we need to find a δ such that

$$\frac{p(1-p)}{n\delta^2} < 0.05. \tag{19.4}$$

But now we're stuck, because p is the parameter that we want to estimate, so how can we do this?

Question: What are some ideas for evaluating (19.4), given we don't know p?

Answer: One idea is to substitute $\overline{X}$ in for p, given that $\overline{X}$ is the estimator for p. However, this only gives us an approximate solution for δ, and we want a guaranteed bound. The idea we use instead is to bound $p(1-p)$.

Question: What is an upper bound on $p(1-p)$?

Answer: $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Thus, from (19.4), we are looking for δ such that

$$\frac{p(1-p)}{n\delta^2} < \frac{1}{4n\delta^2} < 0.05,$$

or equivalently,

$$\delta > \sqrt{\frac{5}{n}}.$$

Notice that this is slightly larger than the value we got in (19.3) via the Chernoff bound, which is to be expected since the Chebyshev bound is weaker than the Chernoff bound and we also upper-bounded the variance. However, like the result in (19.3), we still have the property that the width of the confidence interval shrinks as $\frac{1}{\sqrt{n}}$ as n grows.

19.2.3 Using Tail Bounds to Get Exact Confidence Intervals in General Settings

We now leave polling and return to the general setting of Section 19.1. We have a r.v. X whose mean, $\mathbf{E}[X]$, we are trying to estimate. We are given random i.i.d. samples of X, denoted by $X_1, X_2, \ldots, X_n$. This time we don't know that the X_i's are Bernoulli distributed. In fact, we assume that we know nothing about the distribution of the X_i's, but we do know $\mathbf{Var}(X_i) = \sigma^2$.

Question: How can we derive a 95% confidence interval on $\mathbf{E}[X]$?

Answer: Given that we don't know the distribution of the X_i's, it's hard to imagine how we can use a Chernoff bound. However, we can definitely use the Chebyshev bound. The process is almost identical to that in Section 19.2.2, except that we don't need to bound $\mathbf{Var}(S_n)$. Specifically, we again define

$$S_n = X_1 + X_2 + \cdots + X_n \qquad \text{and} \qquad \overline{X} = \frac{S_n}{n}.$$

Our confidence interval on $\mathbf{E}[X]$ again takes the form

$$\left[\overline{X} - \delta, \overline{X} + \delta\right],$$

where we're seeking δ such that

$$\mathbf{P}\left\{\left|\overline{X} - \mathbf{E}[X]\right| > \delta\right\} < 5\%,$$

or equivalently, such that

$$\mathbf{P}\{|S_n - n\mathbf{E}[X]| > n\delta\} < 5\%.$$

We now use the fact that we know that

$$\mathbf{Var}(S_n) = n\sigma^2$$

to invoke the Chebyshev bound. So we're seeking δ such that

$$\mathbf{P}\left\{|S_n - n\mathbf{E}\left[X\right]| > n\delta\right\} \leq \frac{\mathbf{Var}(S_n)}{n^2\delta^2} = \frac{n\sigma^2}{n^2\delta^2} < 0.05.$$

Solving this, we have that

$$\delta > \frac{\sqrt{20}\sigma}{\sqrt{n}},$$

yielding the confidence interval

$$\left[\overline{X} - \frac{\sqrt{20}\sigma}{\sqrt{n}}\ ,\ \overline{X} + \frac{\sqrt{20}\sigma}{\sqrt{n}}\right], \tag{19.5}$$

where σ refers to σ_{X_i}.

As a final example, we consider how to generate confidence intervals around a signal in a noisy environment.

Example 19.3 (Interval estimation of signal with noise)

Suppose that we're trying to estimate a signal θ (this is a constant), but the signal is sent in a noisy environment where a noise, W, is added to it. The noise, W, has zero mean and variance σ_W^2. We obtain n samples, $X_1, \ldots, X_n$, where

$$X_i = \theta + W_i,$$

and where the W_i's are i.i.d. and $W_i \sim W$.

Again,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

serves as a point estimator for θ.

Question: How can we produce a 95% confidence interval around θ?

Hint: Can we say that the X_i's are i.i.d.?

Answer: The X_i's are in fact i.i.d. Furthermore, $\mathbf{Var}(X_i) = \mathbf{Var}(W)$, which is known. Hence we can directly apply our result from (19.5) to get the following

95% confidence interval for θ:

$$\left[\overline{X} - \frac{\sqrt{20}\sigma_W}{\sqrt{n}}\ ,\ \overline{X} + \frac{\sqrt{20}\sigma_W}{\sqrt{n}}\right]. \tag{19.6}$$

19.3 Approximate Confidence Intervals

In the previous section, we were able to use the Chernoff or Chebyshev bounds to derive guaranteed (exact) confidence intervals in many situations, subject to any desired confidence level. However there are also situations where this is not possible. Furthermore, there are situations where we might *choose* to derive an approximate confidence interval, despite being able to derive an exact confidence interval.

Question: Why would we ever want an approximate confidence interval when we can get an exact one?

Answer: Recall from Chapter 18 that, when the number of samples is high, CLT can offer a much better tail approximation than all existing tail bounds. Thus, even though CLT is just an approximation, we might prefer it to absolute bounds.

As an example of a situation where we might prefer an approximate confidence interval, let's return to the setup in Section 19.2.3. Here, we have a r.v. X whose mean, $\mathbf{E}[X]$, we are trying to estimate. We are given random i.i.d. samples of X, denoted by $X_1, X_2, \ldots, X_n$. All we know about the X_i's is their variance: $\mathbf{Var}(X_i) = \sigma^2$. Our point estimate for $\mathbf{E}[X]$ is

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

which is approximately Normally distributed. Our goal is to derive an interval of the form

$$[\overline{X} - \delta, \overline{X} + \delta],$$

where

$$\mathbf{P}\left\{\left|\overline{X} - \mathbf{E}[X]\right| > \delta\right\} < 5\%.$$

Question: You may recall from Chapter 9 that with probability $\approx 95\%$ the Normal distribution is within 2 standard deviations of its mean.[1] Can we therefore

[1] While it is more precise to write 1.96 standard deviations, we're going with 2 for easy readability.

conclude that an approximate confidence interval for $\mathbf{E}[X]$ is

$$\left[\overline{X} - 2\sigma\ ,\ \overline{X} + 2\sigma\right]?$$

Answer: No, this is wrong. We need to be using $\sigma_{\overline{X}}$ rather than σ, where

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

The derivation of the approximate confidence interval proceeds as usual. Since $\overline{X}$ is a sum of i.i.d. random variables, we can write

$$Q = \frac{\overline{X} - \mathbf{E}[X]}{\sigma_{\overline{X}}} \sim \text{Normal}(0,1), \qquad \text{when } n \to \infty.$$

Hence,

$$\begin{aligned}
\mathbf{P}\{-2 \le Q \le 2\} &\approx 95\% \\
\mathbf{P}\left\{-2 \le \frac{\overline{X} - \mathbf{E}[X]}{\frac{\sigma}{\sqrt{n}}} \le 2\right\} &\approx 95\% \\
\mathbf{P}\left\{-2\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mathbf{E}[X] \le 2\frac{\sigma}{\sqrt{n}}\right\} &\approx 95\% \\
\mathbf{P}\left\{\overline{X} - 2\frac{\sigma}{\sqrt{n}} \le \mathbf{E}[X] \le \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right\} &\approx 95\%.
\end{aligned}$$

Thus, our confidence interval for $\mathbf{E}[X]$ is

$$\left[\overline{X} - 2\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right]. \tag{19.7}$$

Question: How does the confidence interval in (19.7) compare with what we derived earlier in (19.5)?

Answer: Clearly the confidence interval in (19.7) is way tighter, even though it's only an approximation.

Because CLT is so often used for confidence intervals, we summarize our results in Theorem 19.4.

Theorem 19.4 (CLT-based approximate confidence interval) *Let X be a r.v. whose mean, $\mathbf{E}[X]$, we are trying to estimate. We are given n random i.i.d. samples of X, denoted by $X_1, X_2, \ldots, X_n$. All we know about the X_i's is their variance: $\mathbf{Var}(X_i) = \sigma^2$.*
Let

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Let $\Phi(\cdot)$ be the cumulative distribution function (c.d.f.) of the standard Normal, and let

$$\Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}, \qquad \text{i.e.,} \quad z_{\frac{\alpha}{2}} \equiv \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Then,

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\,,\ \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{19.8}$$

is a $(1-\alpha)\cdot 100\%$ approximate confidence interval for $\mathbf{E}[X]$.

We now very briefly turn to the hardest case. Again X is a r.v. whose mean, $\mathbf{E}[X]$, we are trying to estimate. Again we are given n random i.i.d. samples of X, denoted by $X_1, X_2, \ldots, X_n$. However, this time we know *absolutely nothing* about the X_i's. We again wish to determine a $(1-\alpha)\cdot 100\%$ confidence interval around $\mathbf{E}[X]$, but we do *not* know $\mathbf{Var}(X_i) = \sigma^2$, so we cannot directly use (19.8).

If we have an upper bound on $\mathbf{Var}(X_i)$, call it σ^2_{max}, then we can of course substitute σ_{max} in for σ in (19.8). However, if we don't even have a bound on σ, then our best bet is to use the sample standard deviation from (15.5):

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2},$$

yielding the following $(1-\alpha)\cdot 100\%$ confidence interval for $\mathbf{E}[X]$:

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\,,\ \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right]. \tag{19.9}$$

Observe that (19.9) is now an approximation on two fronts. First, we're using CLT, which is an approximation, and second we're approximating $\mathbf{Var}(X_i)$ by the sample variance, S^2. Thus, in using (19.9) it is even more important that n is high.

19.4 Balls and Bins

We now turn to a very different application of tail bounds, illustrated in Figure 19.1, where balls are thrown uniformly at random into bins.

Figure 19.1 *Throwing balls into bins uniformly at random.*

Let's consider the simplest case where we have exactly n balls, each of which is thrown uniformly at random into one of n bins.

Question: On average, how many balls should each bin have?

Answer: Each bin should have one ball in expectation.

Question: What's the highest number of balls that a bin can have?

Answer: n.

This kind of problem comes up in many computer science applications. One example is load balancing of jobs among servers. Each job is routed to a random server, in the hope that all servers end up with an equal number of jobs. The reality, however, is that some servers will end up being sent a lot more jobs than others.

In Exercise 19.8 you will argue that, with high probability, some bin receives $\Omega\left(\frac{\ln n}{\ln \ln n}\right)$ balls. In fact, Exercise 19.7 points out that we expect to have several such "overly full" bins. This says that our attempt at random load balancing is not as "balanced" as we might think.

In Theorem 19.6, we will argue the other side, namely that with high probability no bin will have more than $O\left(\frac{\ln n}{\ln \ln n}\right)$ balls.

Definition 19.5 *The term* **"with high probability" (w.h.p.)** *generally refers to something on the order of* $1 - \frac{1}{n}$*, where* n *is the size of the problem. Sometimes the term is used a little more loosely to refer to something on the order of* $1 - \frac{1}{n^c}$*, where* $c > 0$ *is some constant. When making w.h.p. probabilistic guarantees, it is common to require that* n *is* **"sufficiently large."**

Question: How should we think about $\frac{\ln n}{\ln \ln n}$?

Answer: If we imagine that n is very large, then

$$1 \ll \frac{\ln n}{\ln \ln n} \ll \frac{\ln n}{10^9} \ll \ln n.$$

Theorem 19.6 *If* n *balls are thrown uniformly at random into* n *bins, then, with probability* $\geq 1 - \frac{1}{n}$*, every bin has* $\leq k$ *balls, where*

$$k = \frac{3 \ln n}{\ln \ln n} - 1\,,$$

assuming sufficiently high n*.*

Proof: Our approach will use Chernoff bounds. An alternative approach, not involving Chernoff bounds, is given in Exercise 19.6.

Consider only the jth bin. Let

$$B_j = \sum_{i=1}^{n} X_i = \text{\# balls in bin } j,$$

where

$$X_i = \begin{cases} 1 & \text{if ball } i \text{ goes in bin } j \\ 0 & \text{if ball } i \text{ doesn't go in bin } j \end{cases}.$$

Question: What is the distribution of B_j?

Answer: $B_j \sim \text{Binomial}(n, \frac{1}{n})$, where $\mathbf{E}\left[B_j\right] = 1$.

Question: We want to show that w.h.p. *every* bin has $\leq k$ balls. How can we do this? We'd like to reduce the problem to looking at an individual bin.

Hint: At first this seems complex, because the bins are clearly not independent. But independence is not necessary ...

Hint: We will invoke the *union bound* (Lemma 2.6), which says that for any events E and F,

$$\mathbf{P}\{E \text{ or } F\} \leq \mathbf{P}\{E\} + \mathbf{P}\{F\}.$$

Answer: We want to show that w.h.p. *every* bin has $\leq k$ balls. Equivalently, we want to show:

$$\mathbf{P}\{\text{There exists a bin with } > k \text{ balls}\} < \frac{1}{n}.$$

Equivalently, we want to show:

$$\mathbf{P}\{B_1 > k \text{ or } B_2 > k \text{ or } \cdots \text{ or } B_n > k\} < \frac{1}{n}.$$

But, invoking the union bound, it suffices to show

$$\mathbf{P}\{B_1 > k\} + \mathbf{P}\{B_2 > k\} + \cdots + \mathbf{P}\{B_n > k\} < \frac{1}{n}.$$

Thus it suffices to show that:

$$\mathbf{P}\left\{B_j > k\right\} < \frac{1}{n^2}$$

for every j.

We will now show that:

$$\mathbf{P}\left\{B_j \geq k+1\right\} < \frac{1}{n^2}.$$

Question: Which Chernoff bound on the Binomial should we use: the pretty bound (Theorem 18.4) or the sometimes stronger bound (Theorem 18.6)?

Answer: We observe that k here (which represents δ in Theorem 18.4) grows as $\ln n$, but not as $\Theta(n)$. Hence it's not likely that Theorem 18.4 will give a great bound. If we look at the Chernoff bound given in Theorem 18.6, we see that the ϵ term there is high compared to $\mathbf{E}\left[B_j\right] = 1$. Thus, it is likely that Theorem 18.6 will produce a good bound.

Observing that $\epsilon = k$ and $\mu = 1$ in Theorem 18.6, we have:

$$\mathbf{P}\left\{B_j \geq 1+k\right\} < \frac{e^k}{(1+k)^{(1+k)}}.$$

Hence, to prove that

$$\mathbf{P}\left\{B_j \geq 1+k\right\} < \frac{1}{n^2},$$

it suffices to prove that:

$$\frac{e^k}{(1+k)^{(1+k)}} \leq \frac{1}{n^2}.$$

This latter inequality can be shown to hold by the following argument, which starts by taking logs of both sides:

$$\begin{aligned}
\frac{e^k}{(1+k)^{(1+k)}} &\leq \frac{1}{n^2} \\
&\Updownarrow \\
k - (1+k)\ln(1+k) &\leq -2\ln n \\
&\Updownarrow \\
\frac{3\ln n}{\ln\ln n} - 1 - \frac{3\ln n}{\ln\ln n} \cdot \ln\left(\frac{3\ln n}{\ln\ln n}\right) &\leq -2\ln n \\
&\Updownarrow \\
\frac{3\ln n}{\ln\ln n} - 1 - \frac{3\ln n}{\ln\ln n} \cdot (\ln 3 + \ln\ln n - \ln\ln\ln n) &\leq -2\ln n \\
&\Updownarrow \\
\frac{3}{\ln\ln n} - \frac{1}{\ln n} - \frac{3}{\ln\ln n} \cdot (\ln 3 + \ln\ln n - \ln\ln\ln n) &\leq -2 \\
&\Updownarrow \\
\frac{3}{\ln\ln n} - \frac{1}{\ln n} - \frac{3\ln 3}{\ln\ln n} - 3 + \frac{3\ln\ln\ln n}{\ln\ln n} &\leq -2 \\
&\Updownarrow \\
o(1) + o(1) + o(1) - 3 + o(1) &\leq -2.
\end{aligned}$$

■

Question: Our proof above requires that n is sufficiently large. Where is this needed?

Answer: In the last line of the proof, we state that a bunch of terms are $o(1)$. As explained in Section 1.6, such a statement requires that n is sufficiently large. Specifically, when we say that each term is $o(1)$, we mean that the term approaches 0 for sufficiently high n.

Question: You'll notice that we wrote each of the $o(1)$ terms with a positive sign. Does it matter if the $o(1)$ terms are positive or negative?

Answer: The sign of the $o(1)$ terms here doesn't matter. For high enough n, each $o(1)$ term is arbitrarily close to 0 (see Corollary 1.18). That is, we can think of each term as within 0.00001 of zero, so we don't care whether the terms are positive or negative.

19.5 Remarks on Balls and Bins

There are many more variants of the balls and bins problem, as this paradigm relates to many different computer science applications. For example, one might have m balls and n bins, where $m \neq n$. We will see an example of this when we discuss hashing in Chapter 20. One might have different "colors" of balls, say red balls and blue balls. The "balls" might represent jobs that arrive over time and are dispatched to random servers. One might also have reduced randomness in throwing the balls. For example, in the "power of two choices" version of the balls-and-bins problem, each ball chooses two random bins and then is thrown in the lesser-loaded of these two bins; see [60].

19.6 Exercises

19.1 **Confidence interval warm-up**
You have collected independent samples $X_1, X_2, \ldots, X_{400}$ from some unknown distribution represented by r.v. X. From these samples, you have derived the sample mean and sample variance:

$$\overline{X} = 10 \qquad S^2 = 144.$$

Construct an approximate 99% confidence interval for $\mathbf{E}[X]$.

19.2 **Confidence interval on mean when variance is known**
Suppose we have a r.v. $X \sim \text{Normal}(\mu, \sigma^2)$. Assume that we know σ^2, but we do not know μ. We would like to produce 95% confidence intervals for $\mu = \mathbf{E}[X]$. We have a *small* number n of i.i.d. random samples of X, denoted by $X_1, X_2, \ldots, X_n$. Unfortunately n is small. What is the tightest (least-width) 95% exact confidence interval that we can produce on $\mathbf{E}[X]$?

19.3 **Confidence intervals on vaccine efficacy**
[Proposed by Weina Wang] We are testing a new vaccine, and we want to determine its effectiveness. To do this, we hold a vaccine trial, where we administer the vaccine to all n of the participants. Two weeks later, we check to see the number of infected participants. We model infection as follows:

- With independent *known* probability z, each person will be exposed to the pathogen during the two-week post-vaccination period.
- If person i is exposed, then independently with *unknown* probability p, the vaccine worked and person i will not get sick.
- On the other hand, if the vaccine didn't work, then person i gets measurably sick upon exposure.

- We call z the exposure rate and p the efficacy rate.

Our goal is to estimate the efficacy rate, p. After the two-week period, we check to see whether each person got sick. Let Y_i be an indicator r.v. which is 1 if person i got sick. Let $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

(a) Define the following estimator of p:

$$\hat{p}(Y_1, \ldots, Y_n) = 1 - \frac{\overline{Y}}{z}.$$

(i) Explain the logic behind this estimator.

(ii) Argue that $\hat{p}(Y_1, \ldots, Y_n)$ is an unbiased estimator of p, meaning that $\mathbf{E}[\hat{p}] = p$.

(b) Consider the following interval estimate for p, with $\epsilon = 0.01$:

$$[\hat{p}(Y_1, \ldots, Y_n) - \epsilon\ ,\ \hat{p}(Y_1, \ldots, Y_n) + \epsilon]\ .$$

(i) Using the Chernoff bound, find a study size n which ensures that the confidence level of the interval estimate exceeds 95%, regardless of the value of p.

(ii) Without using the Chernoff bound, find a study size n which ensures that the confidence level of the interval estimate exceeds 95%, regardless of the value of p.

19.4 Interval estimation

[Proposed by Weina Wang] I have a number, $\theta \in (0, 1)$. You don't know θ, but you're allowed to make n guesses $X_1, X_2, \ldots, X_n$. You make your guesses independently and uniformly at random from $(0, 1)$, so $X_i \sim \text{Uniform}(0, 1)$. Your goal is to get within ϵ of θ where ϵ is some specific value in $(0, 1)$. After you make your n guesses, I label those that are "below" θ in blue and those that are "above" in red, as shown in Figure 19.2. Let Y be the largest of the blue $X_1, \ldots, X_n$ (if there are no blue X_i, then $Y = 0$). Let $(Y, Y + \epsilon)$ (yellow interval) be an interval estimate of θ. You would like to be able to say that the interval $(Y, Y + \epsilon)$ contains θ with probability $\geq 1 - \delta$.

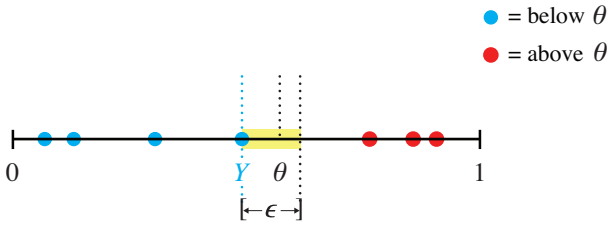

Figure 19.2 *The yellow interval is an interval estimate for Exercise 19.4.*

(a) Compute the c.d.f. of Y, denoted by $F_Y(y)$. Note the range of y.

(b) How large should n be to ensure that $\theta \in (Y, Y+\epsilon)$ with probability $\geq 1-\delta$?

19.5 **Expected size of fullest bin**
In this chapter, we examined throwing n balls uniformly at random at n bins, and we looked at the fullest of the n bins. We proved that with high probability, the fullest bin has $\leq k$ balls, where $k = \frac{3\ln n}{\ln\ln n} - 1$, assuming that n is sufficiently high. Explain why it follows that the expected size of the fullest bin is $O\left(\frac{\ln n}{\ln\ln n}\right)$.

19.6 **High-probability upper bound on number of balls in max bin**
Consider throwing n balls into n bins, uniformly at random. As usual assume that n is sufficiently large. Let $k = \frac{3\ln n}{\ln\ln n}$. In this problem we will prove that the "max bin" (the one with the most balls) has $< k$ balls with high probability. Unlike the chapter, the proof will not use Chernoff bounds. Instead simpler bounds like the union bound will be useful. We will need several helping steps.

(a) First prove the following lemma, which you will need for later steps: If $1 < i < n$, then

$$\binom{n}{i} \leq \left(\frac{ne}{i}\right)^i. \tag{19.10}$$

[Hint: It helps to start by proving that $\binom{n}{i} < \frac{n^i}{i!}$.]

(b) Prove the following lemma, which you will need for later steps:

$$\text{If } k = \frac{3\ln n}{\ln\ln n}, \text{ then } k^k \geq n^{2.99}.$$

[Hint: The argument here resembles that used at the end of this chapter.]

(c) Given that $k = \frac{3\ln n}{\ln\ln n}$, prove that

$$\mathbf{P}\{\text{Bin } j \text{ has } \geq k \text{ balls}\} \leq \frac{1}{n^2}.$$

[Hint: Start by using a union bound over subsets to argue that

$$\mathbf{P}\{\text{Bin } j \text{ has } \geq k \text{ balls}\} \leq \binom{n}{k} \cdot \frac{1}{n^k}.$$

Then use part (a) and then part (b).]

(d) Prove that w.h.p. the maximum bin has $< k$ balls.

19.7 **Lots of bins have lots of balls**
Consider throwing n balls into n bins, uniformly at random. Let $k = \frac{c\ln n}{\ln\ln n}$, where $c = \frac{1}{3}$. Prove that the expected number of bins with at least k balls is $\Omega(n^{2/3})$, for n sufficiently large. We recommend the following steps:

(a) Prove that, for sufficiently high n,

$$\mathbf{P}\{\text{Bin } j \text{ has } \geq k \text{ balls}\} \geq \frac{1}{2ek^k}.$$

[Hint: It will suffice to lower bound the probability that bin j has exactly k balls. You will also use the fact that the function $\left(1-\frac{1}{n}\right)^n$ is increasing with n, and thus exceeds half its limit for high n. It also helps to recall from (1.19) that $\binom{n}{k} > \left(\frac{n}{k}\right)^k$.]

(b) Prove the following lemma, which you will need in the next part:

$$\text{If } k = \frac{c \ln n}{\ln \ln n} \text{ then } k^k \leq n^c.$$

(c) Using parts (a) and (b), show that

$$\mathbf{E}\left[\text{Number of bins with } \geq k \text{ balls}\right] \geq \Omega(n^{1-c}).$$

Specifically, you will show that

$$\mathbf{E}\left[\text{Number of bins with } \geq k \text{ balls}\right] \geq \frac{1}{2e}n^{1-c} = \frac{1}{2e}n^{\frac{2}{3}}.$$

(d) Does part (c) imply that, in expectation, (at least) some constant proportion of the n bins has $\geq k$ balls? For instance, can we conclude that $1/4$ of the bins have $\geq k$ balls, or some other constant fraction?

19.8 **High-probability lower bound on number of balls in max bin**

Consider throwing n balls into n bins, uniformly at random. Let $k = \frac{c \ln n}{\ln \ln n}$, where $c = \frac{1}{3}$. Our goal is to show that with reasonably high probability, at least some bin has $\geq k$ balls.

Let X denote the number of bins with at least k balls. Observe that $X = \sum_{i=1}^{n} X_i$, where X_i is an indicator r.v. equal to 1 if bin i has $\geq k$ balls, and 0 otherwise. We want to prove that

$$\mathbf{P}\{X = 0\} \leq 4e^2 n^{-c} = \frac{4e^2}{n^{\frac{1}{3}}}.$$

(a) Use Chebyshev to upper bound $\mathbf{P}\{X = 0\}$ in terms of $\mathbf{Var}(X)$ and $\mathbf{E}[X]$.

(b) Prove that

$$\mathbf{Var}(X) \leq n.$$

In proving the above, you can assume the following fact (without proof):

$$\mathbf{Var}(X) = \sum_i \mathbf{Var}(X_i) + \sum_{i \neq j} \mathbf{Cov}(X_i, X_j)$$

where

$$\mathbf{Cov}(X_i, X_j) = \mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}\left[X_j\right])\right].$$

The term $\mathbf{Cov}(X_i, X_j)$ stands for "covariance of X_i and X_j," where positive covariance indicates that the random variables are positively correlated and negative covariance indicates that they are negatively correlated. [Hint: As part of your proof, you will need to prove that $\mathbf{Cov}(X_i, X_j) \leq 0$.]

(c) Now use the result from Exercise 19.7(c) and your results from (a) and (b) to finish the proof.

19.9 **Chernoff bound for real-valued random variables**

[Proposed by Vanshika Chowdhary] Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables with values in $[0, 1]$. Assume that $\mathbf{E}[X_i] = \mu_i$. Let

$$X = X_1 + \cdots + X_n.$$

You are given that $\mu = \mathbf{E}[X] \leq 1$ and that $b = \frac{3 \ln n}{\ln \ln n}$.
Show that

$$\mathbf{P}\{X \geq b\} \leq \frac{1}{n^{2.99}}$$

for sufficiently high n. Please follow these steps:

(a) Start with the usual Chernoff bound approach to evaluating $\mathbf{P}\{X \geq b\}$. You will get an expression involving a product of $\mathbf{E}\left[e^{tX_i}\right]$ terms.

(b) Show that $\mathbf{E}\left[e^{tX_i}\right] \leq e^{\mu_i(e^t - 1)}, \quad \forall t > 0$. Here are some helping steps:

(i) Recall from Definition 5.21 that a real-valued function, $g(\cdot)$, defined on interval $S \subseteq \mathbb{R}$ is *convex* if $\forall \lambda \in [0, 1]$, and $\forall \alpha, \beta \in S$,

$$\lambda g(\alpha) + (1 - \lambda) g(\beta) \geq g(\lambda \alpha + (1 - \lambda)\beta).$$

Now use the fact that e^x is a convex function and the fact that $X_i \in [0, 1]$ to show that: $e^{tX_i} \leq X_i e^t + (1 - X_i)e^0$.

(ii) Show that $\mathbf{E}\left[e^{tX_i}\right] \leq e^{\mu_i(e^t - 1)}$.

(c) Substituting the result from (b) into (a), prove $\mathbf{P}\{X \geq b\} \leq e^{b - b \ln b}$.

(d) Now plug in $b = \frac{3 \ln n}{\ln \ln n}$ to get the final result.