1. Large (and moderate) deviation theory
LARGE (AND MODERATE) DEVIATION THEORY
Binomial distribution

Number of heads when a coin with heads bias $p \in [0, 1]$ is tossed $n$ times:

**binomial distribution**

$$S \sim \text{Bin}(n, p)$$
Binomial distribution

Number of heads when a coin with heads bias $p \in [0, 1]$ is tossed $n$ times:

**binomial distribution**

$$S \sim \text{Bin}(n, p)$$

**Basic combinatorics:** for any $k \in \{0, 1, 2, \ldots, n\}$,

$$\Pr[S = k] = \binom{n}{k} p^k (1 - p)^{n-k}.$$
Let $X_1, X_2, \ldots, X_n$ be iid $\text{Bern}(p)$ random variables, and let $S \sim \text{Bin}(n, p)$. Then $S$ has the same distribution as $X_1 + X_2 + \cdots + X_n$. 
Binomial = sums of iid Bernoullis

Let $X_1, X_2, \ldots, X_n$ be iid $\text{Bern}(p)$ random variables, and let $S \sim \text{Bin}(n, p)$. Then $S$ has the same distribution as $X_1 + X_2 + \cdots + X_n$.

**Mean**: By *linearity of expectation*,

$$
\mathbb{E}[S] = \mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np.
$$
Let \( X_1, X_2, \ldots, X_n \) be iid \( \text{Bern}(p) \) random variables, and let \( S \sim \text{Bin}(n, p) \). Then \( S \) has the same distribution as \( X_1 + X_2 + \cdots + X_n \).

**Mean:** By *linearity of expectation*,

\[
\mathbb{E}[S] = \mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np.
\]

**Variance:** Since \( X_1, X_2, \ldots, X_n \) are independent,

\[
\text{var}(S) = \text{var} \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{var}(X_i) = np(1 - p).
\]
**Deviations from the mean**

**Question:** What are the “typical” values (i.e., non-tail event) of $S \sim \text{Bin}(n, p)$?
Deviation from the mean

**Question**: What are the “typical” values (i.e., non-tail event) of $S \sim \text{Bin}(n, p)$?

How do we rigorously quantify the probability mass in the **tails**?
Deviations from the mean

**Question**: What are the “typical” values (i.e., non-tail event) of \( S \sim \text{Bin}(n, p) \)?

How do we rigorously quantify the probability mass in the tails? Differentiate between large and moderate deviations from the mean.
CHERNOFF BOUND: LARGE DEVIATIONS

Let \( S \sim \text{Bin}(n, p) \), and define

\[
\text{RE}(a\|b) := a \ln \frac{a}{b} + (1 - a) \ln \frac{1 - a}{1 - b} \geq 0 \quad (= 0 \text{ iff } a = b)
\]

(relative entropy between Bernoulli distributions with heads biases \( a \) and \( b \)).
Let $S \sim \text{Bin}(n, p)$, and define

$$RE(a\|b) := a \ln \frac{a}{b} + (1 - a) \ln \frac{1 - a}{1 - b} \geq 0 \quad (= 0 \text{ iff } a = b)$$

(relative entropy between Bernoulli distributions with heads biases $a$ and $b$).

**Upper tail bound:** For any $u > p$,

$$\Pr[S \geq n \cdot u] \leq \exp(-n \cdot RE(u\|p)).$$

**Lower tail bound:** For any $\ell < p$,

$$\Pr[S \leq n \cdot \ell] \leq \exp(-n \cdot RE(\ell\|p)).$$
Let \( S \sim \text{Bin}(n, p) \), and define

\[
\text{RE}(a\|b) := a \ln \frac{a}{b} + (1 - a) \ln \frac{1 - a}{1 - b} \geq 0 \quad (= 0 \text{ iff } a = b)
\]

(\textit{relative entropy} between Bernoulli distributions with heads biases \( a \) and \( b \)).

**Upper tail bound**: For any \( u > p \),

\[
\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p)).
\]

**Lower tail bound**: For any \( \ell < p \),

\[
\Pr[S \leq n \cdot \ell] \leq \exp(-n \cdot \text{RE}(\ell\|p)).
\]

Both exponentially small in \( n \).
Chernoff bound: large deviations

Let $S \sim \text{Bin}(n, p)$, and define

$$\text{RE}(a \| b) := a \ln \frac{a}{b} + (1 - a) \ln \frac{1 - a}{1 - b} \geq 0 \quad (= 0 \text{ iff } a = b)$$

(relative entropy between Bernoulli distributions with heads biases $a$ and $b$).

**Upper tail bound:** For any $u > p$,

$$\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u \| p)).$$

**Lower tail bound:** For any $\ell < p$,

$$\Pr[S \leq n \cdot \ell] \leq \exp(-n \cdot \text{RE}(\ell \| p)).$$

Both exponentially small in $n$.

Large deviations from mean $p \cdot n$ (e.g., $(u - p) \cdot n$) are exponentially unlikely.
ILLUSTRATION OF LARGE DEVIATIONS

\[ p = \frac{1}{3}, \quad u = \frac{1}{3} + 0.05, \quad n = 100 \]

\[ \exp(-\text{RE}(u||p)) \approx 0.995 \]
Illustration of large deviations

\[ p = \frac{1}{3}, \quad u = \frac{1}{3} + 0.05, \quad n = 200 \]

\[ \exp(-\text{RE}(u\|p)) \approx 0.995 \]
Illustration of large deviations

\[ p = \frac{1}{3}, \quad u = \frac{1}{3} + 0.05, \quad n = 300 \]

\[ \exp(-\text{RE}(u \| p)) \approx 0.995 \]
Illustration of large deviations

\[ p = 1/3, \quad u = 1/3 + 0.05, \quad n = 400 \]

\[ \exp(-\text{RE}(u\|p)) \approx 0.995 \]
$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 500$

$\exp(-\text{RE}(u\|p)) \approx 0.995$
Illustration of large deviations

\[ p = \frac{1}{3}, \quad u = \frac{1}{3} + 0.05, \quad n = 600 \]

\[ \exp(-\text{RE}(u\|p)) \approx 0.995 \]
Illustration of Large Deviations

$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 700$

$\exp(-\text{RE}(u||p)) \approx 0.995$
Illustration of large deviations

\[ p = \frac{1}{3}, \quad u = \frac{1}{3} + 0.05, \quad n = 800 \]

\[ \exp(-\text{RE}(u||p)) \approx 0.995 \]
Illustration of large deviations

\[ p = \frac{1}{3}, \quad u = \frac{1}{3} + 0.05, \quad n = 900 \]
\[ \exp(-\text{RE}(u\|p)) \approx 0.995 \]
Illustration of large deviations

\[ p = \frac{1}{3}, \quad u = \frac{1}{3} + 0.05, \quad n = 1000 \]

\[ \exp(-\text{RE}(u\|p)) \approx 0.995 \]
**Proof of Chernoff bound (upper tail bound)**

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$. 

- Consider $n$ iid Bernoulli random variables: $X_1, X_2, \cdots, X_n$. 
- Let $E \subseteq \{0, 1\}^n$ be all outcomes $x = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$. 

Some shorthand notation:

- $p[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $p$. 
- $u[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $u$. 

Core of the proof: Consider any outcome $x \in E$ with, say, $k \geq n \cdot u$ heads:

\[
 p[x] u[x] = p[k](1-p)^{n-k} u[k](1-u)^{n-k} \leq (pu)^k(1-p-1-u)^n \cdot (1-u).
\] 

\[
 \Pr[S \geq n \cdot u] = \sum_{x \in E} p[x] \leq \sum_{x \in E} u[x] (pu)^n \cdot (1-p-1-u)^n \cdot (1-u) \leq \exp(-n \cdot \text{RE}(u\|p)).
\]
**Proof of Chernoff Bound (Upper Tail Bound)**

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \ldots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $x = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$. 

Some shorthand notation:

- $p[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $p$.
- $u[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $u$.

The core of the proof: Consider any outcome $x \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads: $p[x] = p^k (1 - p)^{n-k} \leq (p u)^k (1 - p - u)^{n-k}$. Then,

$$
\Pr[S \geq n \cdot u] = \sum_{x \in \mathcal{E}} p[x] \leq \sum_{x \in \mathcal{E}} u[x] \leq (p u)^n (1 - p - u)^n (1 - u) \leq \exp(-n \cdot \text{RE}(u\|p)).
$$
**Proof of Chernoff Bound (Upper Tail Bound)**

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u \| p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \ldots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $x = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

**Some shorthand notation:**

- $p[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $p$.
- $u[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $u$. 
Proof of Chernoff Bound (Upper Tail Bound)

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \cdots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

**Some shorthand notation:**
- $p[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $p$.
- $u[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $u$.

**Core of the proof:** Consider any outcome $\mathbf{x} \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:
Proof of Chernoff Bound (Upper Tail Bound)

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \cdots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

**Some shorthand notation:**
- $p[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $p$.
- $u[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $u$.

**Core of the proof:** Consider any outcome $\mathbf{x} \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

\[
\frac{p[\mathbf{x}]}{u[\mathbf{x}]} \]
**Proof of Chernoff Bound (Upper Tail Bound)**

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u \| p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \ldots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $x = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

**Some shorthand notation:**
- $p[x] := \text{probability mass of outcome } x \text{ when } X_i \text{ has heads bias } p$.
- $u[x] := \text{probability mass of outcome } x \text{ when } X_i \text{ has heads bias } u$.

**Core of the proof:** Consider any outcome $x \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

$$\frac{p[x]}{u[x]} = \frac{p^k (1 - p)^{n-k}}{u^k (1 - u)^{n-k}}.$$
**Proof of Chernoff bound (upper tail bound)**

**Theorem**: For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \ldots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

**Some shorthand notation:**

- $p[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $p$.
- $u[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $u$.

**Core of the proof**: Consider any outcome $\mathbf{x} \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

\[
\frac{p[\mathbf{x}]}{u[\mathbf{x}]} = \frac{p^k (1 - p)^{n-k}}{u^k (1 - u)^{n-k}} = \left(\frac{p}{u}\right)^k \left(\frac{1 - p}{1 - u}\right)^{n-k}
\]
**Proof of Chernoff Bound (Upper Tail Bound)**

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \ldots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $x = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

**Some shorthand notation:**

- $p[x] := \text{probability mass of outcome } x \text{ when } X_i \text{ has heads bias } p$.
- $u[x] := \text{probability mass of outcome } x \text{ when } X_i \text{ has heads bias } u$.

**Core of the proof:** Consider any outcome $x \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

$$\frac{p[x]}{u[x]} = \frac{p^k(1-p)^{n-k}}{u^k(1-u)^{n-k}} = \left(\frac{p}{u}\right)^k \left(\frac{1-p}{1-u}\right)^{n-k} \leq \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1-p}{1-u}\right)^{n \cdot (1-u)}.$$
Proof of Chernoff bound (upper tail bound)

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \cdots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $x = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

**Some shorthand notation:**

- $p[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $p$.
- $u[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $u$.

**Core of the proof:** Consider any outcome $x \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

$$\frac{p[x]}{u[x]} = \frac{p^k (1-p)^{n-k}}{u^k (1-u)^{n-k}} = \left(\frac{p}{u}\right)^k \left(\frac{1-p}{1-u}\right)^{n-k} \leq \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1-p}{1-u}\right)^{n \cdot (1-u)}.$$

$$\Pr[S \geq n \cdot u] = \sum_{x \in \mathcal{E}} p[x]$$
Proof of Chernoff bound (Upper tail bound)

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u||p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \cdots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $x = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

**Some shorthand notation:**
- $p[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $p$.
- $u[x] :=$ probability mass of outcome $x$ when $X_i$ has heads bias $u$.

**Core of the proof:** Consider any outcome $x \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

$$
\frac{p[x]}{u[x]} = \frac{p^k (1-p)^{n-k}}{u^k (1-u)^{n-k}} = \left(\frac{p}{u}\right)^k \left(\frac{1-p}{1-u}\right)^{n-k} \leq \left(\frac{p}{u}\right)^{n\cdot u} \left(\frac{1-p}{1-u}\right)^{n\cdot(1-u)}.
$$

$$
\Pr[S \geq n \cdot u] = \sum_{x \in \mathcal{E}} p[x] \leq \sum_{x \in \mathcal{E}} u[x] \left(\frac{p}{u}\right)^{n\cdot u} \left(\frac{1-p}{1-u}\right)^{n\cdot(1-u)}
$$
**Proof of Chernoff Bound (Upper Tail Bound)**

**Theorem:** For $S \sim \text{Bin}(n,p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u||p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \cdots, X_n$. Let $\mathcal{E} \subseteq \{0,1\}^n$ be all outcomes $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

**Some shorthand notation:**
- $p[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $p$.
- $u[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $u$.

**Core of the proof:** Consider any outcome $\mathbf{x} \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

$$\frac{p[\mathbf{x}]}{u[\mathbf{x}]} = \frac{p^k(1 - p)^{n-k}}{u^k(1 - u)^{n-k}} = \left(\frac{p}{u}\right)^k \left(\frac{1 - p}{1 - u}\right)^{n-k} \leq \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1 - p}{1 - u}\right)^{n \cdot (1-u)}.$$

$$\Pr[S \geq n \cdot u] = \sum_{\mathbf{x} \in \mathcal{E}} p[\mathbf{x}] \leq \sum_{\mathbf{x} \in \mathcal{E}} u[\mathbf{x}] \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1 - p}{1 - u}\right)^{n \cdot (1-u)} \leq \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1 - p}{1 - u}\right)^{n \cdot (1-u)}.$$
Proof of Chernoff bound (upper tail bound)

**Theorem:** For $S \sim \text{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p))$ for $u > p$.

Consider $n$ iid Bernoulli random variables: $X_1, X_2, \cdots, X_n$. Let $\mathcal{E} \subseteq \{0, 1\}^n$ be all outcomes $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

**Some shorthand notation:**
- $p[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $p$.
- $u[\mathbf{x}] :=$ probability mass of outcome $\mathbf{x}$ when $X_i$ has heads bias $u$.

**Core of the proof:** Consider any outcome $\mathbf{x} \in \mathcal{E}$ with, say, $k \geq n \cdot u$ heads:

$$
\frac{p[\mathbf{x}]}{u[\mathbf{x}]} = \frac{p^k (1 - p)^{n-k}}{u^k (1 - u)^{n-k}} = \left(\frac{p}{u}\right)^k \left(\frac{1 - p}{1 - u}\right)^{n-k} \leq \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1 - p}{1 - u}\right)^{n \cdot (1 - u)}
$$

$$
\Pr[S \geq n \cdot u] = \sum_{\mathbf{x} \in \mathcal{E}} p[\mathbf{x}] \leq \sum_{\mathbf{x} \in \mathcal{E}} u[\mathbf{x}] \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1 - p}{1 - u}\right)^{n \cdot (1 - u)} \leq \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1 - p}{1 - u}\right)^{n \cdot (1 - u)} = \exp(-n \cdot \text{RE}(u\|p)).
$$
Moderate deviations

What about more moderate deviations of size $o(n)$?
What about more moderate deviations of size $o(n)$?

"Fact": $S \sim \text{Bin}(n, p)$ “typically” in $[np - 2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}]$. 
What about more moderate deviations of size $o(n)$?

"Fact": $S \sim \text{Bin}(n, p)$ “typically” in $[np - 2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}]$.

$\text{Bin}(10, 1/3)$

$np \approx 3.333$, $2\sqrt{np(1-p)} \approx 2.9814$
What about more moderate deviations of size $o(n)$?

**“Fact”**: $S \sim \text{Bin}(n, p)$ “typically” in $[np - 2\sqrt{np(1 - p)}, np + 2\sqrt{np(1 - p)}]$.

$\operatorname{Bin}(100, 1/3)$

$np \approx 33.333, \quad 2\sqrt{np(1 - p)} \approx 9.4281$
What about more moderate deviations of size $o(n)$?

"Fact": $S \sim \text{Bin}(n, p)$ “typically” in $[np - 2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}]$.
To rigorously quantify moderate deviations, can again use Chernoff bound

\[ \Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p)), \]

but ask how small can \( u \) be before the bound exceeds some fixed \( \delta \in (0, 1) \)?
To **rigorously quantify moderate deviations**, can again use Chernoff bound

\[
\Pr[S \ge n \cdot u] \le \exp(-n \cdot \text{RE}(u||p)),
\]

but ask how small can \(u\) be before the bound exceeds some fixed \(\delta \in (0, 1)\)?

By calculus, for \(u > p\),

\[
\text{RE}(u||p) \ge \frac{(u - p)^2}{2u}.
\]

Therefore, for \(u > p\),

\[
\Pr[S \ge n \cdot u] \le \exp(-n \cdot \text{RE}(u||p)) \le \exp\left(-n \cdot \frac{(u - p)^2}{2u}\right).
\]
To **rigorously quantify moderate deviations**, can again use Chernoff bound

\[
\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p)),
\]

but ask how small can \( u \) be before the bound exceeds some fixed \( \delta \in (0, 1) \)?

By calculus, for \( u > p \),

\[
\text{RE}(u\|p) \geq \frac{(u - p)^2}{2u}.
\]

Therefore, for \( u > p \),

\[
\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \text{RE}(u\|p)) \leq \exp\left(-n \cdot \frac{(u - p)^2}{2u}\right).
\]

By algebra, the RHS is \( \leq \delta \) when

\[
n \cdot u = n \cdot p + \sqrt{2np \ln(1/\delta)} + 2\ln(1/\delta) = n \cdot p + O(\sqrt{n}).
\]
Similar argument for lower tail.
Similar argument for lower tail.

By calculus, for \( \ell < p \leq 1/2 \),

\[
\text{RE}(\ell\|p) \geq \frac{(p - \ell)^2}{2p}.
\]

Therefore, for \( \ell < p \leq 1/2 \),

\[
\Pr[S \leq n \cdot \ell] \leq \exp(-n \cdot \text{RE}(\ell\|p)) \leq \exp\left(-n \cdot \frac{(p - \ell)^2}{2p}\right).
\]

By algebra, the RHS is \( \delta \) when

\[
n \cdot \ell = n \cdot p - \sqrt{2np \ln(1/\delta)} = n \cdot p - O(\sqrt{n}).
\]
**Moderate deviations**

**Combining upper and lower tail bounds:** for $p \leq 1/2$,

$$\Pr\left \{ S \in \left [ np - \sqrt{2np \ln(1/\delta)} , np + \sqrt{2np \ln(1/\delta) + 2 \ln(1/\delta)} \right ] \right \} \geq 1 - 2\delta.$$

**Union bound:** $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$. 

1-2$\delta$
**Moderate deviations**

**Combining upper and lower tail bounds:** for $p \leq 1/2$,

$$\Pr\left\{ S \in \left[ np - \sqrt{2np \ln(1/\delta)}, \; np + \sqrt{2np \ln(1/\delta) + 2 \ln(1/\delta)} \right] \right\} \geq 1 - 2\delta.$$ 

**Union bound:** \( \Pr[A \cup B] \leq \Pr[A] + \Pr[B] \)

Approximately recovers previous “fact” that \( S \) is “typically” in \( \left[ np - 2\sqrt{np(1-p)}, \; np + 2\sqrt{np(1-p)} \right] \) (though a bit looser).
**Estimating a Coin Bias**

**Another interpretation:** estimating heads bias \( p \leq 1/2 \) from iid sample \( X_1, X_2, \ldots, X_n \) with

\[
\hat{p} := \frac{X_1 + X_2 + \cdots + X_n}{n}.
\]

With probability at least \( 1 - 2\delta \),

\[
p - \sqrt{\frac{2p \ln(1/\delta)}{n}} \leq \hat{p} \leq p + \sqrt{\frac{2p \ln(1/\delta)}{n}} + \frac{2 \ln(1/\delta)}{n};
\]

i.e., the estimate \( \hat{p} \) is usually reasonably close to the truth \( p \).
Another interpretation: estimating heads bias $p \leq 1/2$ from iid sample $X_1, X_2, \ldots, X_n$ with

$$\hat{p} := \frac{X_1 + X_2 + \cdots + X_n}{n}. $$

With probability at least $1 - 2\delta$,

$$p - \sqrt{\frac{2p \ln(1/\delta)}{n}} \leq \hat{p} \leq p + \sqrt{\frac{2p \ln(1/\delta)}{n}} + \frac{2 \ln(1/\delta)}{n};$$

i.e., the estimate $\hat{p}$ is usually reasonably close to the truth $p$.

How close? Depends on:

- whether you’re asking about how far above $p$ or how far below $p$ (upper and lower tails are somewhat asymmetric);
- the sample size $n$;
- the true heads bias $p$ itself;
- the “confidence” parameter $\delta$. 
Let $\hat{f} : \mathcal{X} \to \mathcal{Y}$ be a classifier, and suppose you have iid test data $T$ (that are independent of $\hat{f}$).
Let $\hat{f}: \mathcal{X} \to \mathcal{Y}$ be a classifier, and suppose you have iid test data $T$ (that are independent of $\hat{f}$).

**True error:**

$$\text{err}(\hat{f}) = \Pr[\hat{f}(X) \neq Y].$$

**Test error:**

$$\text{err}(\hat{f}, T) = \frac{1}{|T|} \sum_{(x,y) \in T} \mathbb{1}\{\hat{f}(x) \neq y\}.$$

**Distribution of test error:**

$$|T| \cdot \text{err}(\hat{f}, T) \sim \text{Bin}(|T|, \text{err}(\hat{f})).$$
APPLICATION: TEST ERROR

Let \( \hat{f}: \mathcal{X} \to \mathcal{Y} \) be a classifier, and suppose you have iid test data \( T \) (that are independent of \( \hat{f} \)).

**True error:**

\[
err(\hat{f}) = \Pr[\hat{f}(X) \neq Y].
\]

**Test error:**

\[
err(\hat{f},T) = \frac{1}{|T|} \sum_{(x,y) \in T} 1\{\hat{f}(x) \neq y\}.
\]

**Distribution of test error:**

\[
|T| \cdot err(\hat{f},T) \sim \text{Bin}(|T|, err(\hat{f})).
\]

**Applying Chernoff bounds:** with prob. \( \geq 1 - 2\delta \) (w.r.t. random draw of \( T \)),

\[
|err(\hat{f}) - err(\hat{f},T)| \leq \sqrt{\frac{2 \text{err}(\hat{f}) \ln(1/\delta)}{|T|}} + \frac{2 \ln(1/\delta)}{|T|}.
\]
Application: Test Error

Let \( \hat{f}: \mathcal{X} \rightarrow \mathcal{Y} \) be a classifier, and suppose you have iid test data \( T \) (that are independent of \( \hat{f} \)).

**True error:**
\[
\text{err}(\hat{f}) = \Pr[\hat{f}(X) \neq Y].
\]

**Test error:**
\[
\text{err}(\hat{f}, T) = \frac{1}{|T|} \sum_{(x, y) \in T} 1\{\hat{f}(x) \neq y\}.
\]

**Distribution of test error:**
\[
|T| \cdot \text{err}(\hat{f}, T) \sim \text{Bin}(|T|, \text{err}(\hat{f})).
\]

**Applying Chernoff bounds:** with prob. \( \geq 1 - 2\delta \) (w.r.t. random draw of \( T \)),
\[
\left| \text{err}(\hat{f}) - \text{err}(\hat{f}, T) \right| \leq \sqrt{\frac{2 \text{err}(\hat{f}) \ln(1/\delta)}{|T|}} + \frac{2 \ln(1/\delta)}{|T|}.
\]

Suggests (very) rough idea of the resolution at which you can distinguish classifiers’ test errors, based on size of test set.
(Estimate of heads bias with $\hat{p} = (X_1 + \cdots + X_n)/n$.)

With probability at least $1 - 2\delta$,

$$p \in \left[ \hat{p} - \sqrt{\frac{2p \ln(1/\delta)}{n}} - \frac{2 \ln(1/\delta)}{n}, \hat{p} + \sqrt{\frac{2p \ln(1/\delta)}{n}} \right].$$
Application: Confidence Intervals

(Estimate of heads bias with $\hat{p} = (X_1 + \cdots + X_n)/n$.)

With probability at least $1 - 2\delta$,

$$p \in \left[\hat{p} - \sqrt{\frac{2p \ln(1/\delta)}{n}} - \frac{2 \ln(1/\delta)}{n}, \hat{p} + \sqrt{\frac{2p \ln(1/\delta)}{n}}\right].$$

Unfortunately interval also depends on $p$. 
(Estimate of heads bias with $\hat{p} = (X_1 + \cdots + X_n)/n$.)

With probability at least $1 - 2\delta$,

$$p \in \left[ \hat{p} - \sqrt{\frac{2p \ln(1/\delta)}{n}} - \frac{2 \ln(1/\delta)}{n}, \hat{p} + \sqrt{\frac{2p \ln(1/\delta)}{n}} \right].$$

**Unfortunately interval also depends on $p$.**

**Fix:** can “solve” for the largest value of $q \in [0, 1]$ such that

$$q \leq \hat{p} + \sqrt{\frac{2q \ln(1/\delta)}{n}}$$

$\longrightarrow$ Upper limit of confidence interval. (Can similarly get lower limit.)
(Estimate of heads bias with $\hat{p} = (X_1 + \cdots + X_n)/n$.)

With probability at least $1 - 2\delta$,

$$p \in \left[\hat{p} - \sqrt{\frac{2p \ln(1/\delta)}{n}} - \frac{2 \ln(1/\delta)}{n}, \hat{p} + \sqrt{\frac{2p \ln(1/\delta)}{n}}\right].$$

Unfortunately interval also depends on $p$.

**Fix:** can “solve” for the largest value of $q \in [0, 1]$ such that

$$q \leq \hat{p} + \sqrt{\frac{2q \ln(1/\delta)}{n}}$$

$\rightarrow$ Upper limit of confidence interval. (Can similarly get lower limit.)

After some more algebra, get confidence intervals in terms of $\hat{p}$:

$$p \in \left[\hat{p} - \sqrt{\frac{2\hat{p} \ln(1/\delta)}{n}} - \frac{2 \ln(1/\delta)}{n}, \hat{p} + \sqrt{\frac{2\hat{p} \ln(1/\delta)}{n}} + \frac{2 \ln(1/\delta)}{n}\right].$$
Sums of iid Bernoulli random variables:

- Large deviations from mean of size $\Omega(n)$ are exponentially unlikely.
- Bulk of probability mass is within moderate deviations of size $O(\sqrt{n})$.
- Applies in many other cases besides sums of iid Bernoulli.

Tool: Chernoff bound

- Reason about test error.
- Construct confidence intervals.