COMS 4771 Lecture 11

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

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

Binomial distribution

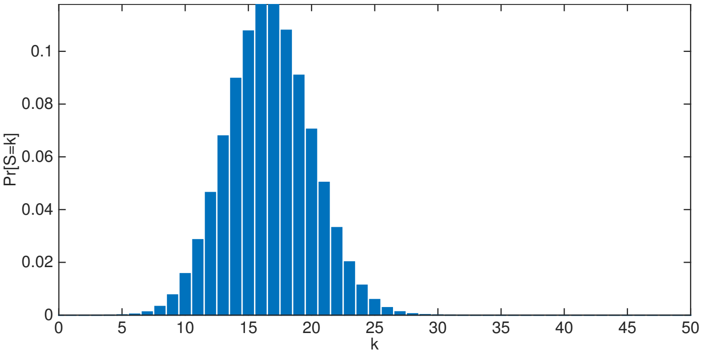
Number of heads when a coin with heads bias $p \in [0, 1]$ is tossed n times:

binomial distribution

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

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

$$\Pr[S = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$



Binomial = sums of iid Bernoullis

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

BINOMIAL = SUMS OF IID BERNOULLIS

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

Mean: By *linearity of expectation*,

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

Binomial = sums of iid Bernoullis

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

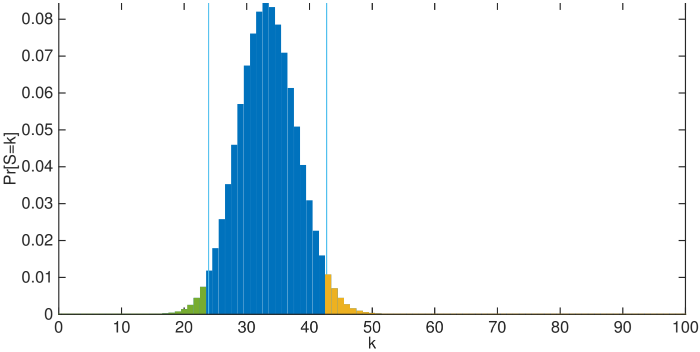
Mean: By *linearity of expectation*,

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

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

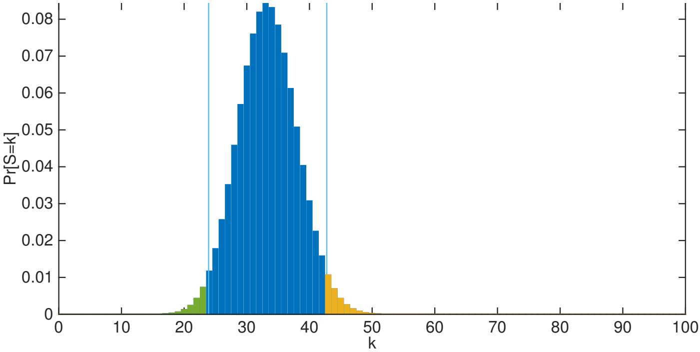
$$\mathrm{var}(S) = \mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{var}(X_i) = np(1-p).$$

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



Deviations from the mean

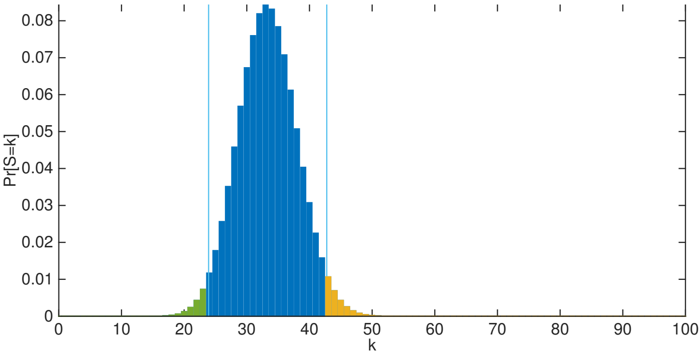
Question: What are the "typical" values (i.e., non-tail event) of $S \sim \mathrm{Bin}(n, p)$?



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



How do we rigorously quantify the probability $\boxed{\text{mass in the } \textbf{tails}\text{?}}$
Differentiate between **large** and **moderate** deviations from the mean.

Chernoff bound: large deviations

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

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

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

Chernoff bound: large deviations

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

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

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

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

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

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

$$\Pr[S \leq n \cdot \ell] \;\leq\; \exp(-n \cdot \mathrm{RE}(\ell, p)).$$

Chernoff bound: large deviations

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

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

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

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

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

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

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

Both exponentially small in n.

use the comparion!!!

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

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

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

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

$$\boxed{\Pr[S \geq n \cdot u] \;\leq}\; \exp(-n \boxed{\mathrm{RE}(u,p)).}$$

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

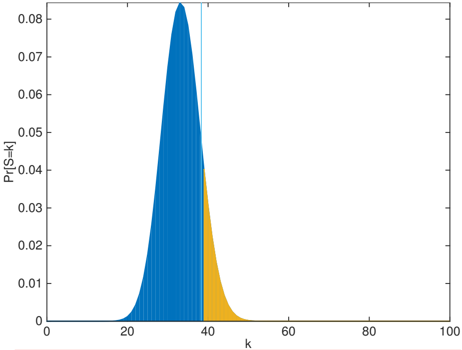
$$\Pr[S \leq n \cdot \ell] \;\leq\; \exp(-n \cdot \mathrm{RE}(\ell,p)).$$

get p's!!!

Both exponentially small in n.

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

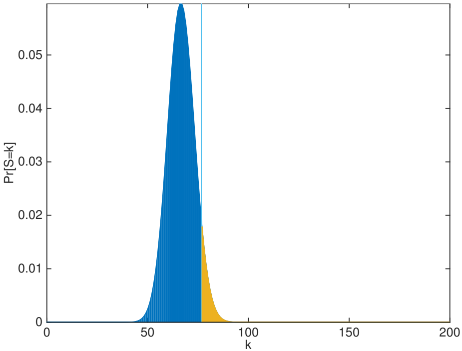
Illustration of large deviations



n = 1
P[S > u]

$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 100$$
$$\exp(-\mathrm{RE}(u,p)) \approx 0.995$$

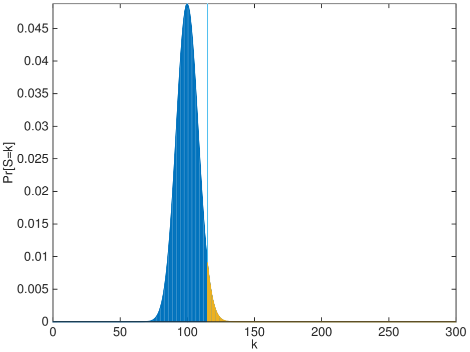
Illustration of large deviations



$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 200$$
$$\exp(-\,\mathrm{RE}(u, p)) \approx 0.995$$

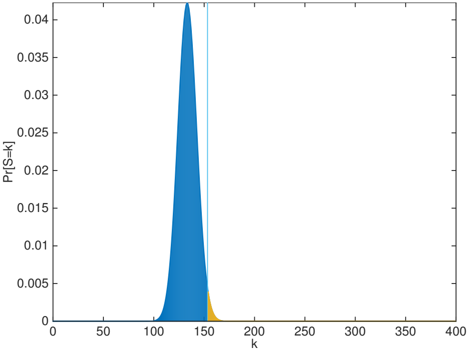
the n would not affect exp value

Illustration of large deviations



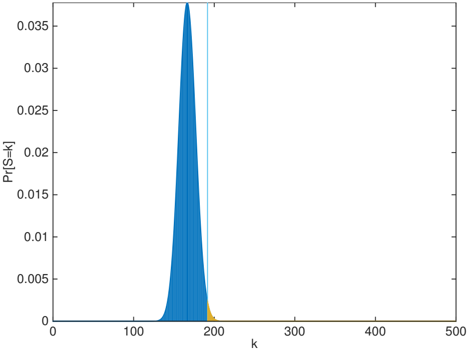
$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 300$$
$$\exp(-\,\mathrm{RE}(u,p)) \approx 0.995$$

Illustration of large deviations



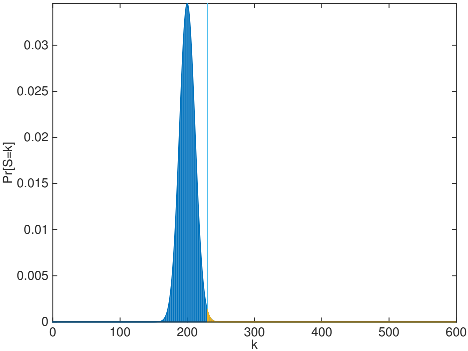
$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 400$$
$$\exp(-\,\mathrm{RE}(u,p)) \approx 0.995$$

Illustration of large deviations



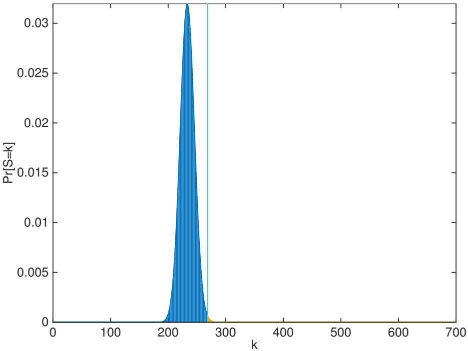
$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 500$$
$$\exp(-\,\mathrm{RE}(u,p)) \approx 0.995$$

Illustration of large deviations



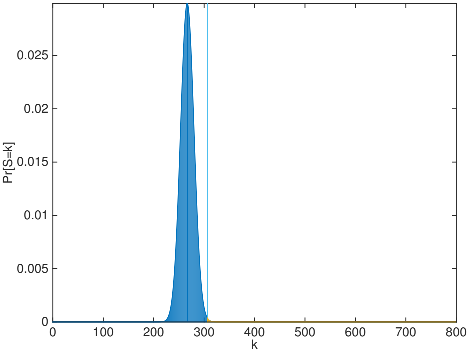
$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 600$$
$$\exp(-\mathrm{RE}(u,p)) \approx 0.995$$

Illustration of large deviations



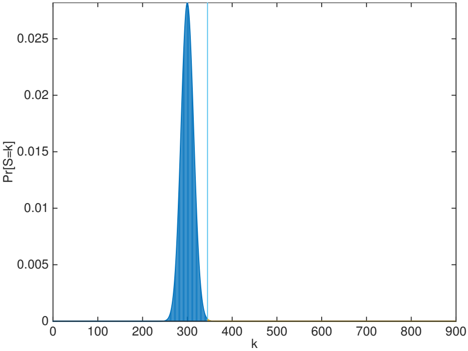
$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 700$$
$$\exp(-\mathrm{RE}(u,p)) \approx 0.995$$

Illustration of large deviations



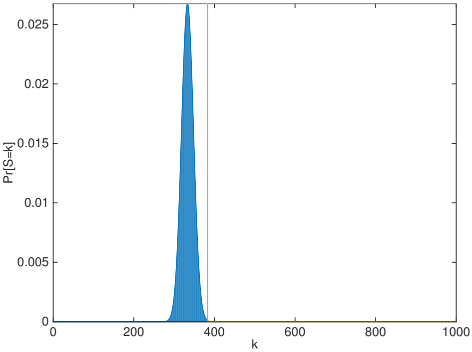
$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 800$$
$$\exp(-\mathrm{RE}(u,p)) \approx 0.995$$

Illustration of large deviations



$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 900$$
$$\exp(-\mathrm{RE}(u, p)) \approx 0.995$$

Illustration of large deviations



$$p = 1/3, \quad u = 1/3 + 0.05, \quad n = 1000$$
$$\exp(-\,\mathrm{RE}(u,p)) \approx 0.995$$

Proof of Chernoff bound (upper tail bound)

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

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n,p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u,p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

Some shorthand notation:

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

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

Some shorthand notation:

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

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

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^{n} x_i \geq n \cdot u$.

Some shorthand notation:

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

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

$$\frac{p[\boldsymbol{x}]}{u[\boldsymbol{x}]}$$

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

Some shorthand notation:

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

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

$$\frac{p[\boldsymbol{x}]}{u[\boldsymbol{x}]} = \frac{p^k(1-p)^{n-k}}{u^k(1-u)^{n-k}}$$

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

Some shorthand notation:

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

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

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

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

Some shorthand notation:

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

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

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

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

Some shorthand notation:

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

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

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

$$\Pr[S \geq n \cdot u] \quad = \quad \sum_{\boldsymbol{x} \in \mathcal{E}} p[\boldsymbol{x}]$$

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

Some shorthand notation:

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

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

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

$$\Pr[S \geq n \cdot u] \quad = \quad \sum_{\boldsymbol{x} \in \mathcal{E}} p[\boldsymbol{x}] \quad \leq \quad \sum_{\boldsymbol{x} \in \mathcal{E}} u[\boldsymbol{x}] \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1-p}{1-u}\right)^{n \cdot (1-u)}$$

Proof of Chernoff bound (upper tail bound)

Theorem: For $S \sim \mathrm{Bin}(n, p)$, $\Pr[S \geq n \cdot u] \leq \exp(-n \cdot \mathrm{RE}(u, p))$ 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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $\sum_{i=1}^n x_i \geq n \cdot u$.

Some shorthand notation:

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

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

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

$$\begin{aligned} \Pr[S \geq n \cdot u] \quad &= \quad \sum_{\boldsymbol{x} \in \mathcal{E}} p[\boldsymbol{x}] \quad \leq \quad \sum_{\boldsymbol{x} \in \mathcal{E}} u[\boldsymbol{x}] \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1-p}{1-u}\right)^{n \cdot (1-u)} \\ &\leq \quad \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1-p}{1-u}\right)^{n \cdot (1-u)} \end{aligned}$$

Proof of Chernoff bound (upper tail bound)

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

all outcomes meet S >= n * u

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

Some shorthand notation:

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

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

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

$$\boxed{\Pr[S \geq n \cdot u] = \sum_{\boldsymbol{x} \in \mathcal{E}} p[\boldsymbol{x}]} \leq \sum_{\boldsymbol{x} \in \mathcal{E}} u[\boldsymbol{x}] \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1-p}{1-u}\right)^{n \cdot (1-u)}$$

$$\leq \left(\frac{p}{u}\right)^{n \cdot u} \left(\frac{1-p}{1-u}\right)^{n \cdot (1-u)} = \exp(-n \cdot \mathrm{RE}(u,p)). \quad \square$$

Moderate deviations

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

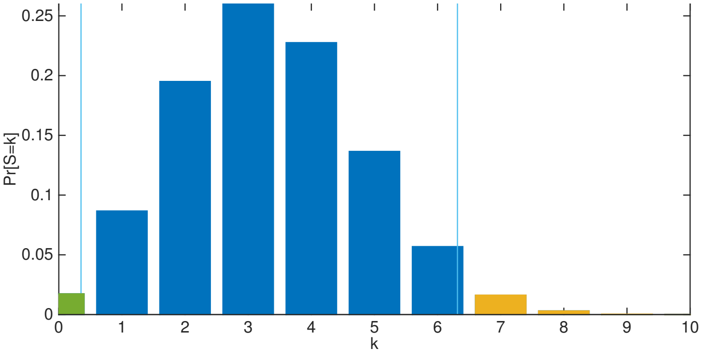
Moderate deviations

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

"**Fact**": $S \sim \mathrm{Bin}(n, p)$ "typically" in $\left[np - 2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}\right]$.

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

"**Fact**": $S \sim \mathrm{Bin}(n, p)$ "typically" in $\left[np - 2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}\right]$.



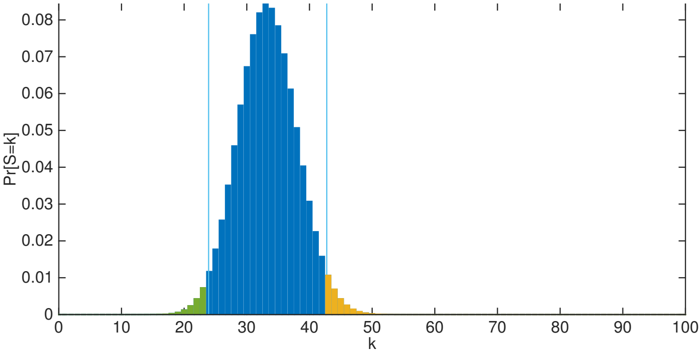
$\mathrm{Bin}(10, 1/3)$

$np \approx 3.333, \;\; 2\sqrt{np(1-p)} \approx 2.9814$

MODERATE DEVIATIONS

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

"**Fact**": $S \sim \mathrm{Bin}(n,p)$ "typically" in $\left[np - 2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}\right]$.



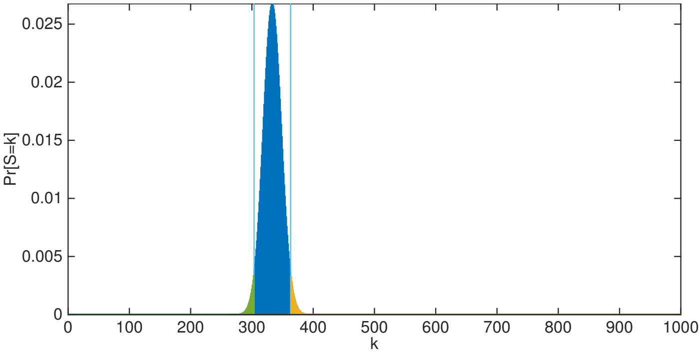
$$\mathrm{Bin}(100, 1/3)$$

$$np \approx 33.333, \;\; 2\sqrt{np(1-p)} \approx 9.4281$$

Moderate deviations

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

"**Fact**": $S \sim \mathrm{Bin}(n, p)$ "typically" in $\left[np - 2\sqrt{np(1-p)}, np + 2\sqrt{np(1-p)}\right]$.



$$\mathrm{Bin}(1000, 1/3)$$
$$np \approx 333.333, \quad 2\sqrt{np(1-p)} \approx 29.8142$$

To **rigorously quantify moderate deviations**, can again use Chernoff bound

$$\Pr[S \geq n \cdot u] \;\leq\; \exp(-n \cdot \mathrm{RE}(u, p)),$$

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

Moderate deviations

To **rigorously quantify moderate deviations**, can again use Chernoff bound

$$\Pr[S \geq n \cdot u] \;\leq\; \exp(-n \cdot \mathrm{RE}(u, p)),$$

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

By calculus, for $u > p$,

$$\mathrm{RE}(u, p) \geq \frac{(u-p)^2}{2u}.$$

Therefore, for $u > p$,

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

Moderate deviations

To **rigorously quantify moderate deviations**, can again use Chernoff bound

$$\Pr[S \geq n \cdot u] \;\leq\; \exp(-n \cdot \mathrm{RE}(u, p)),$$

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

By calculus, for $u > p$,

$$\mathrm{RE}(u, p) \geq \frac{(u-p)^2}{2u}.$$

Therefore, for $u > p$,

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

By algebra, the RHS is δ 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.

Moderate deviations

Similar argument for lower tail.

By calculus, for $\ell < p \le 1/2$,

$$\mathrm{RE}(\ell, p) \ge \frac{(p-\ell)^2}{2p}.$$

Therefore, for $\ell < p \le 1/2$,

$$\Pr[S \le n \cdot \ell] \;\le\; \exp(-n \cdot \mathrm{RE}(\ell, p)) \;\le\; \exp\left(-n \cdot \frac{(p-\ell)^2}{2p}\right).$$

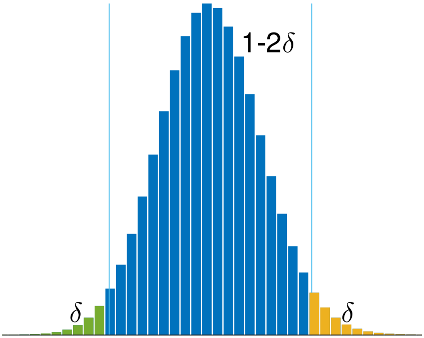
By algebra, the RHS is δ when

$$n \cdot \ell \;=\; n \cdot p - \sqrt{2np\ln(1/\delta)} \;=\; n \cdot p - O(\sqrt{n}).$$

Moderate deviations

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

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

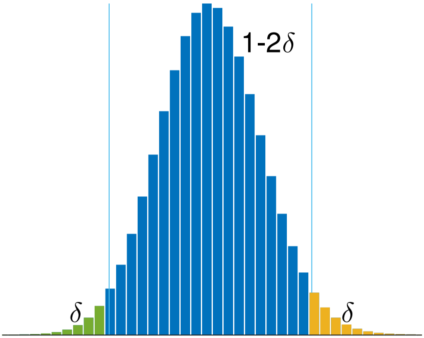


Union bound: $\Pr[A \cup B] \le \Pr[A] + \Pr[B]$

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.

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.

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 δ.

Application: test error

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

Application: test error

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

True error:

for a classifier

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

Test error:

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

Distribution of test error:

$$|T| \cdot \operatorname{err}(\hat{f}, T) \sim \operatorname{Bin}(|T|, \operatorname{err}(\hat{f})).$$

Application: test error

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

True error:

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

Test error:

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

Distribution of test error:

$$|T| \cdot \operatorname{err}(\hat{f}, T) \sim \operatorname{Bin}(|T|, \operatorname{err}(\hat{f})).$$

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

$$\left|\operatorname{err}(\hat{f}) - \operatorname{err}(\hat{f}, T)\right| \;\leq\; \sqrt{\frac{2\operatorname{err}(\hat{f})\ln(1/\delta)}{|T|}} + \frac{2\ln(1/\delta)}{|T|}.$$

don't use test result to adjust model

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:

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

Test error:

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

At the test size grow. the Gap between true error and test error is minimized!

Distribution of test error:

$$\boxed{|T| \cdot \operatorname{err}(\hat{f}, T)} \sim \operatorname{Bin}(|T|, \operatorname{err}(\hat{f})).$$

even the same X, could lead to many value

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

$$\left|\operatorname{err}(\hat{f}) - \operatorname{err}(\hat{f}, T)\right| \;\leq\; \sqrt{\frac{2 \operatorname{err}(\hat{f}) \ln(1/\delta)}{|T|}} + \frac{2 \ln(1/\delta)}{|T|}.$$

Suggests (very) **rough idea** of the resolution at which you can distinguish classifiers' test errors, based on size of test set.

Application: confidence intervals

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

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

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

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.

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.

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.)

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.

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.)

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

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

Summary and final remarks

- 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.