Binomial distribution

Binomial distribution

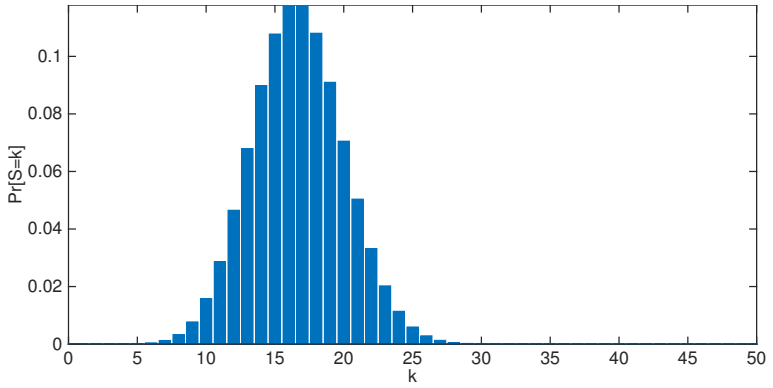
Number of heads when a coin with heads bias $p \in [0,1]$ is tossed n times:

binomial distribution

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

Probability mass function: for any $k \in \{0,1,2,\ldots,n\}$,

$$\mathbb{P}(S=k) = \binom{n}{k} p^k (1-p)^{n-k}\,.$$



Special case: Bernoulli distribution

The outcome of a coin toss with heads bias $p \in [0,1]$:

Bernoulli distribution

$$X \sim \mathrm{Bern}(p) = \mathrm{Bin}(1,p)$$

$$\mathbb{P}(X=1) = p\,, \quad \mathbb{P}(X=0) = 1-p\,.$$

Mean:

$$\mathbb{E}(X) = \mathbb{P}(X=0)\cdot 0 + \mathbb{P}(X=1)\cdot 1 = p\,.$$

Variance:

$$\mathrm{var}(X) = \mathbb{E}\Big[\big(X-\mathbb{E}(X)\big)^2\Big] = p(1-p)\,.$$

(Standard deviation is $\sqrt{\mathrm{var}(X)}$; more convenient to use than $\mathbb{E}\big[|X-\mathbb{E}(X)|\big]$.)

Binomial = sums of i.i.d. Bernoullis

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\mathrm{Bern}(p)$ 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*,

$$\mathbb{E}(S) = \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) = np\,.$$

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

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

Test error rate

Let $\hat{f}\colon \mathcal{X} \to \mathcal{Y}$ be a classifier, and suppose you have i.i.d. test data T (that are *independent of* $\hat{f}$); let $n := |T|$.

True error rate (with $(X, Y) \sim \mathbb{P}$):

$$\mathrm{err}(\hat{f}) \;=\; \mathbb{P}(\hat{f}(X) \neq Y)\,.$$

Test error rate:

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

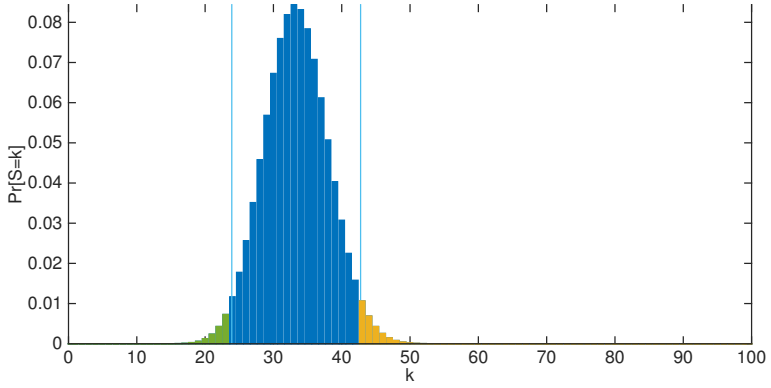
The random variables $\{\mathbb{1}\{\hat{f}(x) \neq y\}\}_{(x,y)\in T}$ are *independent* and *identically distributed* as $\mathrm{Bern}(\mathrm{err}(\hat{f}))$.

Distribution of test error rate:

$$n \cdot \mathrm{err}(\hat{f}, T) \;\sim\; \mathrm{Bin}(n, \mathrm{err}(\hat{f}))\,.$$

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 quantify the probability mass in the **tails**?

Chernoff bound: large deviations

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

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

the *relative entropy* between Bernoulli distributions with heads biases a and b. (Measures how different the distributions are.)

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

$$\mathbb{P}(S \geq n \cdot u) \;\leq\; \exp\big(-\mathrm{RE}(u\|p) \cdot n\big)\,.$$

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

$$\mathbb{P}(S \leq n \cdot \ell) \;\leq\; \exp\big(-\mathrm{RE}(\ell\|p) \cdot n\big)\,.$$

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

Consider $S \sim \mathrm{Bin}(n, 1/3)$ and $u = 1/3 + 0.05 \approx 0.383$.

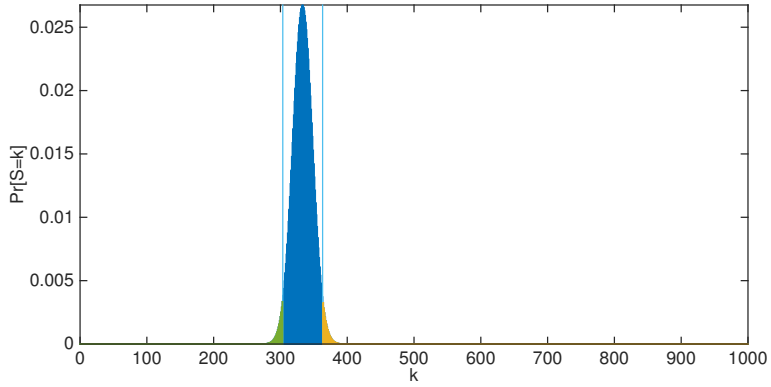
$$\exp(-\mathrm{RE}(u\|p)) \approx 0.995$$

What is $\mathbb{P}(S \geq u)$?

Moderate deviations

How large are "typical" deviations?

"**Fact**": $S \sim \mathrm{Bin}(n,p)$ "typically" within few standard deviations from mean.



$\mathrm{Bin}(1000, 1/3)$

$np \approx 333.333, \;\; 2\sqrt{np(1-p)} \approx 29.8142$

Moderate deviations

To derive the "fact', can again use Chernoff bound

$$\mathbb{P}(S \geq n \cdot u) \;\leq\; \exp\bigl(-\,\mathrm{RE}(u\|p) \cdot n\bigr)\,.$$

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 $u > p$,

$$\mathbb{P}(S \geq n \cdot u) \;\leq\; \exp\bigl(-\,\mathrm{RE}(u\|p) \cdot n\bigr) \;\leq\; \exp\left(-\frac{(u-p)^2}{2u} \cdot n\right).$$

By algebra, the RHS is δ when

$$n \cdot u \;=\; n \cdot p + \sqrt{2np\ln(1/\delta)} + 2\ln(1/\delta)\,.$$

Moderate deviations

Similar argument for lower tail.

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

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

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

$$\mathbb{P}(S \leq n \cdot \ell) \;\leq\; \exp\bigl(-\,\mathrm{RE}(\ell\|p) \cdot n\bigr) \;\leq\; \exp\left(-\frac{(p-\ell)^2}{2p} \cdot n\right).$$

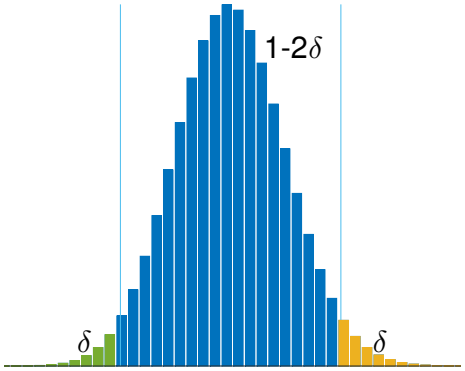
By algebra, the RHS is δ when

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

Moderate deviations

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

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



Union bound: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

Estimating a coin bias

Another interpretation: estimating heads bias $p \leq 1/2$ from i.i.d. sample $X_1, X_2, \ldots, X_n$ with

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

So

$$\mathbb{P}\left(p - \sqrt{\frac{2p\ln(1/\delta)}{n}} \;\leq\; \hat{p} \;\leq\; p + \sqrt{\frac{2p\ln(1/\delta)}{n}} + \frac{2\ln(1/\delta)}{n}\right) \;\geq\; 1 - 2\delta\,;$$

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.

Suggests **rough idea** of the resolution at which you can distinguish classifiers' error rates, based on size of test set.

Key takeaways

1. Large ($\Omega(n)$) and "typical" ($O(\sqrt{n})$) deviations for $\mathrm{Bin}(n, p)$.
2. Use of Chernoff bound to reason about error rates.