Informative projections → } next quarter
Singular value decomposition → }

Sampling

DSE 210

Data (e.g. one particular digit) in $\mathbb{R}^{784}$

~ 6000 data pts

$$\begin{bmatrix} 1000 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}$$

Σ : 784×784

very often, this will be singular (or very very close to singular)

In order to compute the density at a point x, we need

$$x^T \Sigma^{-1} x$$

$\hat{\Sigma}$: emplrical covariance matrix

may be singular

$\hat{\Sigma} + cI$ is not singular for any $c > 0$

$\hat{\Sigma} = np.cov(X, \_)$ ← bias correction

then

$\hat{\Sigma} \leftarrow \hat{\Sigma} + cI$

① How to choose c?

② Why not just make it tiny, since that's enough to make $\hat{\Sigma}$ invertible?

① Basic idea: try a bunch of values, pick the one that's "best." accuracy on validation set.

- typically narrow the range of values by first trying exponentially-spaced options eg. 10, 100, 1000
- useful to first get an idea of the rough scale by looking at the average of the diagonal entries of $\hat{\Sigma}$.

$$\hat{\Sigma} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.01 \end{bmatrix}$$

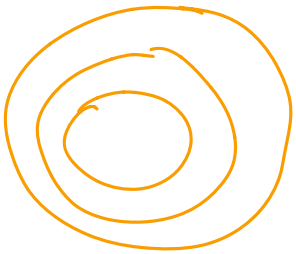
$$\hat{\Sigma} = \begin{bmatrix} 100 & 0 \\ 0 & 10 \end{bmatrix}$$

② Σ is a big matrix relative to the amt of data.

Adding cI is basically a smoothing operation that compensates for a shortage of data.

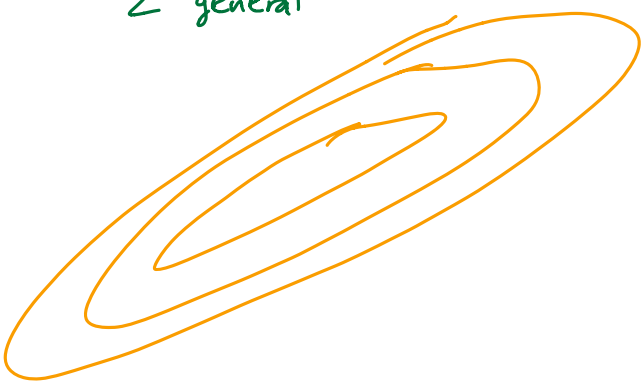
Gaussian pictures

Σ = identity or $c \cdot I$

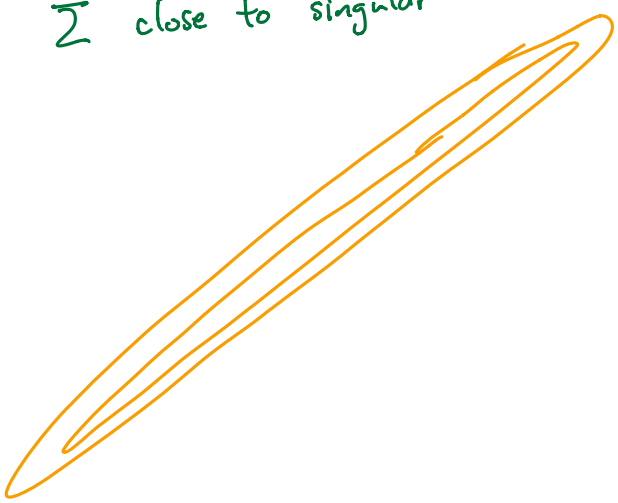


"spherical"

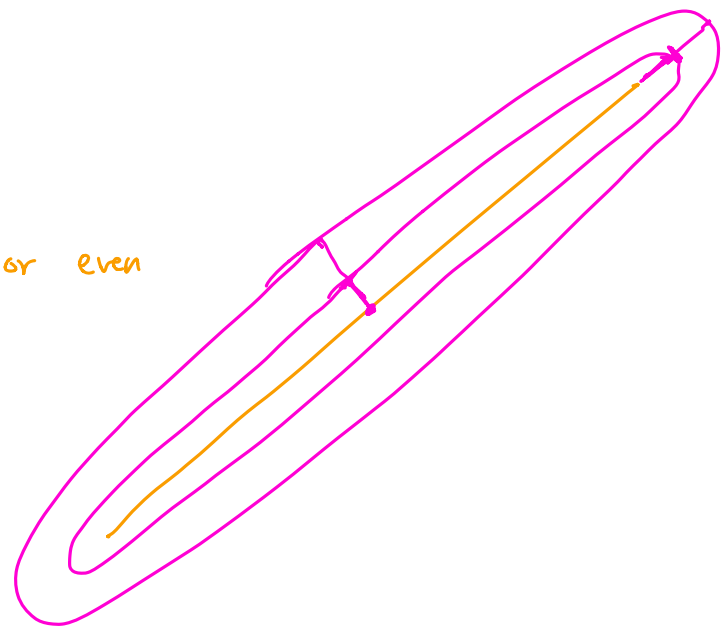
Σ general



Σ close to singular



or even



effect of adding cI to Σ

MNIST

① Load data (70,000 pts = 60,000 training
10,000 test)

② Fit a multivariate Gaussian to each digit

using np.mean
np.cov

③ Finding a value of c using validation set

④ Evaluating on test set of 10000 pts

10000 pts × 10 Gaussians : 100,000 log pdf calculations

$x^T \left(\Sigma_j^{-1}\right) x$

---

① At test time, ignore words that were not in the training set.

② Downweight words that are common across classes

- "stopwords" – removed altogether
- other common words – downweight

Outline

1. Laws of large numbers
2. Basic sampling designs
3. Confidence intervals

Review: Expected value

The expected value of a random variable X is

$$\mathbb{E}(X) = \sum_x x \Pr(X = x).$$

Linearity properties:

- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ for any random variable X and any constants a, b.
- $\mathbb{E}(X_1 + \cdots + X_k) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_k)$ for any random variables $X_1, X_2, \ldots, X_k$.

Example: Toss n coins of bias p (heads probability), and let X be the number of heads (in the range 0 to n). What is $\mathbb{E}(X)$?

Let the individual coin tosses be $X_1, \ldots, X_n \in \{0,1\}$ (0 = tails, 1 = heads).

- $\mathbb{E}[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$
- $X = X_1 + \cdots + X_n$

  $\therefore \mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] = np$

Review: Variance

Maximizing $p(1-p)$ (calculus): $f(p) = p - p^2$, $f'(p) = 1 - 2p = 0 \Rightarrow p = 1/2$

Variance of an r.v. X is $\text{var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2) - \mu^2$, where $\mu = \mathbb{E}(X)$.

Useful variance rules:

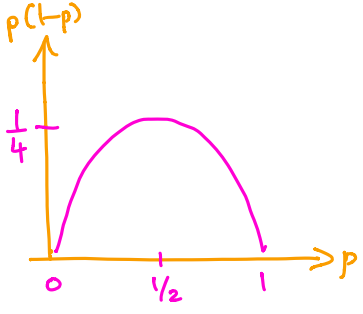
- $\text{var}(X_1 + \cdots + X_k) = \text{var}(X_1) + \cdots + \text{var}(X_k)$ if X_i's independent.
- $\text{var}(aX + b) = a^2\text{var}(X)$.

The standard deviation of X is $\sqrt{\text{var}(X)}$. It is (an approximation to) the average amount by which X differs from its mean.

tails, heads

Example: Toss a coin of bias p. Let $X \in \{0, 1\}$ be the outcome. What are the variance and standard deviation of X? For what p is the variance highest?

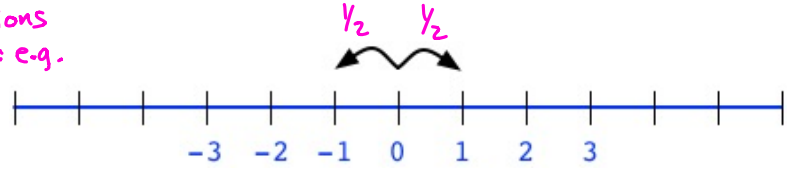
- $\mathbb{E}[X] = p$
- $\mathbb{E}[X^2] = \mathbb{E}[X] = p$ since $X^2 = X$ (because $0^2 = 0$ and $1^2 = 1$)
- $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p)$
- $\text{std}(X) = \sqrt{p(1-p)}$ ← maximized at $p = \frac{1}{2}$

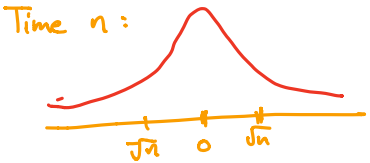


Variance of a sum

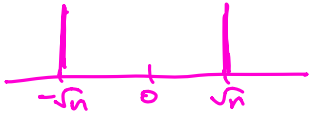
We are interested in: X_n his position after n steps.

Symmetric random walk. A drunken man sets out from a bar. At each time step, he either moves one step to the right or one step to the left, with equal probabilities. Roughly where is he after n steps?





There are many distributions with mean 0 and std $\sqrt{n}$: e.g.



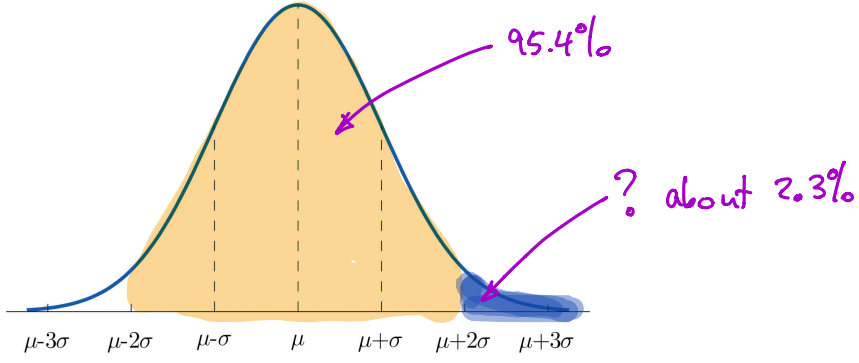
Let $X_i \in \{-1, 1\}$ be his ith step. His position after n steps is $X = X_1 + \cdots + X_n$

- What are $\mathbb{E}(X_i)$ and $\text{var}(X_i)$? $\mathbb{E}[X_i] = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0$; $\mathbb{E}[X_i^2] = 1$; $\text{var}(X_i) = 1$
- What are $\mathbb{E}(X)$ and $\text{var}(X)$?
  $\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = 0$; $\text{var}(X) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n$
- What is $\text{std}(X)$?
  $\text{std}(X) = \sqrt{n}$
- What is the distribution over his possible positions?

As we will see, it is (roughly) a Gaussian distribution with mean 0 and variance n.

The normal distribution

Symmetric distribution



The normal (or *Gaussian*) $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- 68.3% of the distribution lies in the range $\mu \pm \sigma$
- 95.4% lies within $\mu \pm 2\sigma$
- 99.7% lies within $\mu \pm 3\sigma$

The central limit theorem : this is the basis of sampling

Suppose $X_1, \ldots, X_n$ are independent, each with mean μ and variance σ^2. } $E[X_i] = \mu$, $\text{var}(X_i) = \sigma^2$

coin flips, movements of drunken man, etc.

Let $S_n = X_1 + \cdots + X_n$. What are the mean and variance of S_n?

Mean: $E[S_n] = n\mu$

Variance: $\text{Var}(S_n) = n\sigma^2$

Central limit theorem, very roughly: For reasonably large n, the distribution of $S_n = X_1 + \cdots + X_n$ looks like $N(n\mu, n\sigma^2)$, the Gaussian with mean $n\mu$ and variance $n\sigma^2$.

Question: What does this imply about the distribution of the **average** $(X_1 + \cdots + X_n)/n$?

Let $A_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}$

$E[A_n] = \frac{E[S_n]}{n} = \mu$

$\text{var}(A_n) = \frac{\text{var}(S_n)}{n^2} = \frac{\sigma^2}{n}$

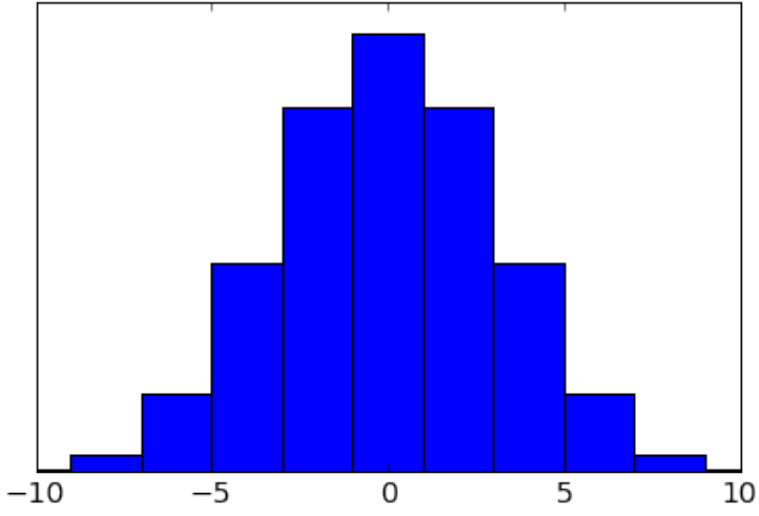
} ∴ Distribution of the average A_n looks like $N(\mu, \frac{\sigma^2}{n})$

the more samples (n), the smaller the variance

Symmetric random walk, again

mean 0
variance n

Each X_i is either 1 or -1, with probability $1/2$. Thus $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



10 steps

Symmetric random walk, again

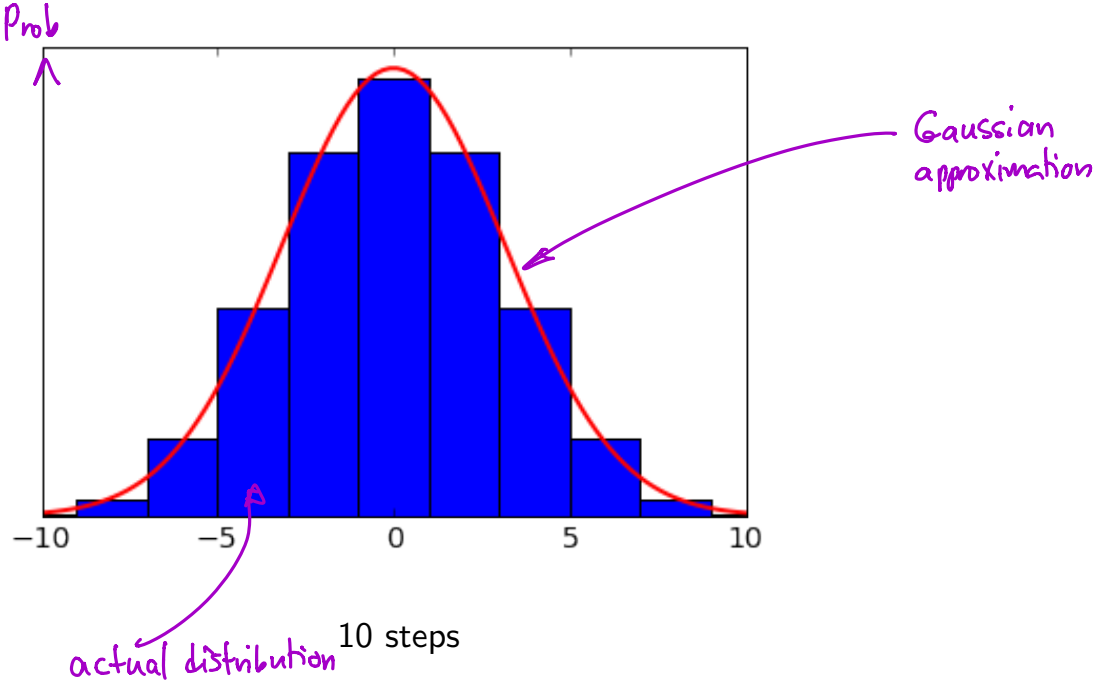
Each X_i is either 1 or -1, with probability $1/2$. Thus $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.

10 steps

$Pr(X = 0)$

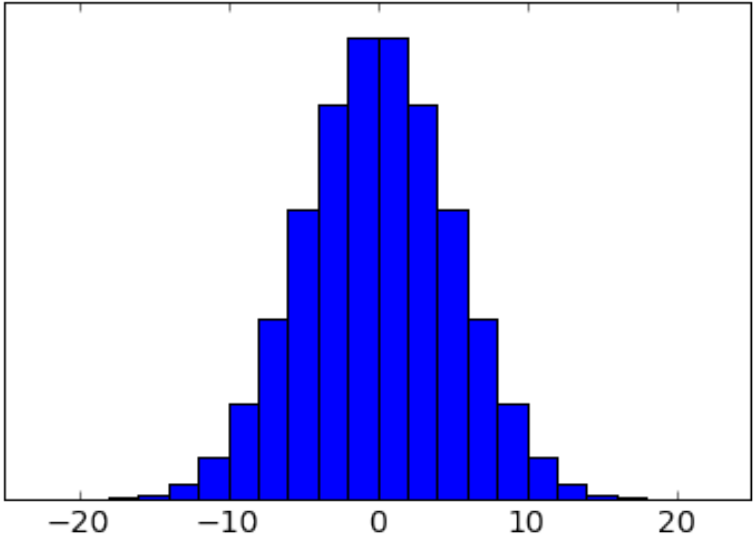
$= Pr(5 \text{ left}, 5 \text{ right})$

$= \binom{10}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5$



Symmetric random walk, again

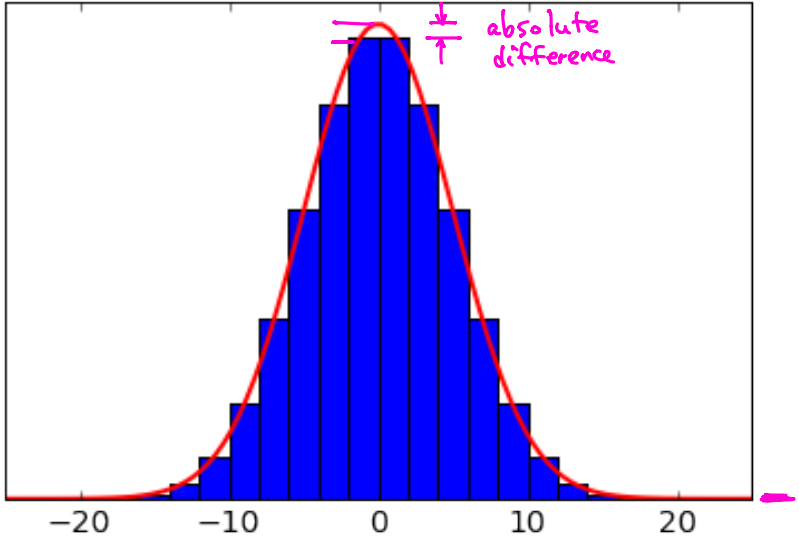
Each X_i is either 1 or -1, with probability $1/2$. Thus $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



25 steps

Symmetric random walk, again

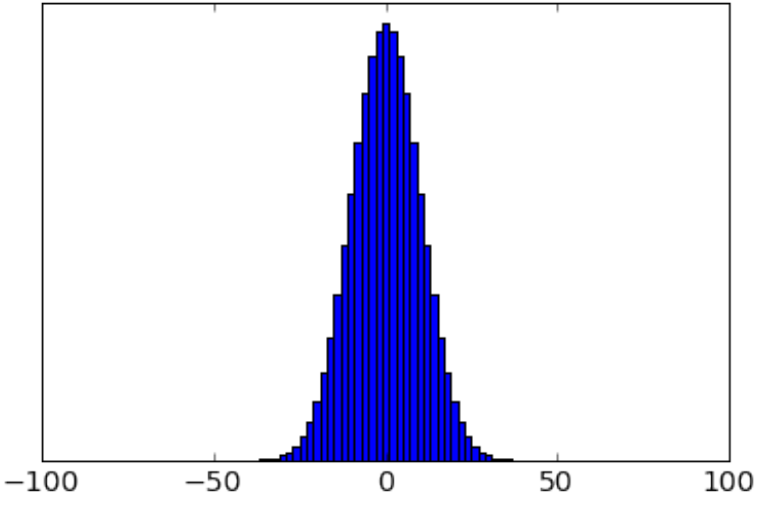
Each X_i is either 1 or -1, with probability $1/2$. Thus $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



25 steps

Symmetric random walk, again

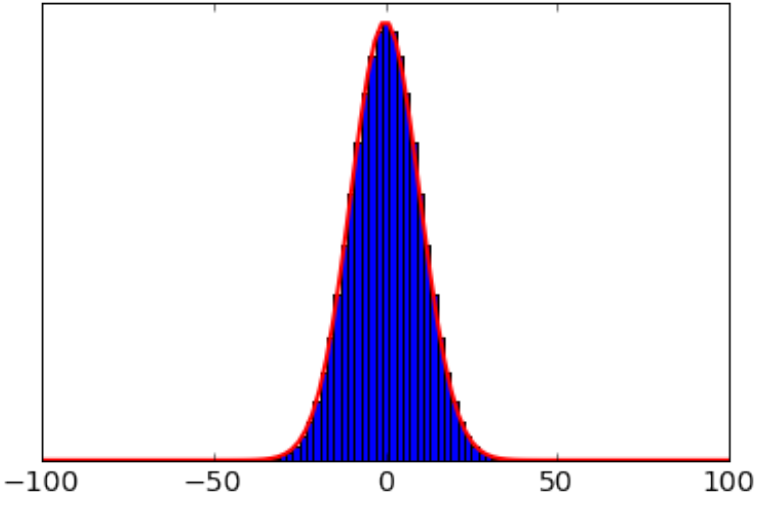
Each X_i is either 1 or -1, with probability $1/2$. Thus $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



100 steps

Symmetric random walk, again

Each X_i is either 1 or -1, with probability $1/2$. Thus $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



100 steps

Tosses of a biased coin

A coin of bias (heads probability) p is tossed n times.

- What is the distribution of the observed **number** of heads, roughly?

Tosses of a biased coin

A coin of bias (heads probability) p is tossed n times.

- What is the distribution of the observed **number** of heads, roughly?
  Answer: $N(np, np(1-p))$
  Mean np, standard deviation on the order of $\sqrt{n}$.
- What is the distribution of the observed **fraction** of heads, roughly?

Tosses of a biased coin

A coin of bias (heads probability) p is tossed n times.

range $0, 1, \ldots, n$

- What is the distribution of the observed **number** of heads, roughly?
  Answer: $N(np, np(1-p))$
  Mean np, standard deviation on the order of $\sqrt{n}$.

SUM

range $[0,1]$

- What is the distribution of the observed **fraction** of heads, roughly?
  Answer: $N(p, p(1-p)/n)$.
  Mean p, standard deviation on the order of $1/\sqrt{n}$.

AVERAGE

Pick a random person in town. Heads = Democrat, Tails = not Democrat.

$p = \frac{12000}{30000} = 0.4$

Example: A town has 30,000 registered voters, of whom 12,000 are Democrats. A random sample of 1,000 voters is chosen. How many of them would we expect to be Democrats, roughly?

$n = 1000$

Number of Democrats will follow a $N(np, np(1-p)) = N(400, 240)$ distribution.

This has mean 400 and std. deviation $\sqrt{240} \simeq 15.5$.

$400 \pm 2\sigma = 360 - 440$

- Toss a fair coin 10,000 times.

  What is a reasonable range for the number of heads you would see?

- X = # heads $\in \{0, 1, 2, \ldots, 10000\}$ $\quad n = 10000$, $p = \frac{1}{2}$

  $\mathbb{E}[X] = 5000 \quad [= np]$

  $\text{var}(X) = 10000 \times \frac{1}{2} \times \frac{1}{2} \quad [= np(1-p)]$

  $\text{std}(X) = 50$

- X has (approximately) a Gaussian distribution, $N(5000, 2500)$

Two standard deviations:

$X \in [4900, 5100]$ w.p. $\geq 95\%$

$X \in [4850, 5150]$ w.p. $\geq 99.7\%$

"99.7% confidence interval for X"

Outline

1. Laws of large numbers
2. Basic sampling designs
3. Confidence intervals

Sampling design

In the 1948 Presidential election, the polls all predicted Thomas Dewey as the winner, with at least a five-point margin. But the outcome was quite different.

Selection bias

The Republican bias in the Gallup Poll, 1936-1948.

Year	Gallup's prediction of Republican vote	Actual Republican vote
1936	44	38
1940	48	45
1944	48	46
1948	50	45

Selection bias

The Republican bias in the Gallup Poll, 1936-1948.

Year	Gallup's prediction of Republican vote	Actual Republican vote
1936	44	38
1940	48	45
1944	48	46
1948	50	45

The safest way to sample is **at random**.

Multistage cluster sampling

Sometimes random sampling is inconvenient, and careful multistage procedures are used.

For instance:

1. Stage 1
   - Divide the US into four geographical regions: Northeast, South, Midwest, West.
   - Within each region, group together all population centers of similar sizes. E.g. All towns in the northeast with 50-250 thousand people.
   - Pick a random sample of these towns.
2. Stage 2
   - Divide each town into wards, and each ward into precincts.
   - Select some wards at random from the towns chosen earlier.
   - Select some precincts at random from among these wards.
   - Then select households at random from these precincts.
   - Then select members of the selected households at random, within the designated age ranges.

Sample size versus population size

A certain town in Illinois has the same balance of Democrats and Republicans as the nation at large. We want to determine these fractions using a random sample of 1000 people. Would it be better to choose the 1000 people from the town in Illinois, or from the entire country?

Let p be the (unknown) fraction of Democrats in this town (and in the country as a whole).

What is the observed fraction of Democrats in random sample of 1000 people from the town?

$n = 1000 \rightarrow N\left(p, \frac{p(1-p)}{1000}\right)$

If we pick 1000 people from the entire country, the distribution is the same.

What matters is the sample size and not the overall population size.

Outline

1. Laws of large numbers
2. Basic sampling designs
3. Confidence intervals

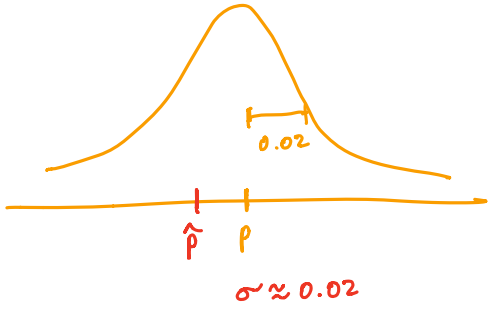
Example: Estimating a fraction

A university has 25,000 registered students. In a survey, 400 students were chosen at random, and it turned out that 317 of them were living at home. Estimate the fraction of students living at home.

Let p be the actual fraction of students living at home.

1. What is the observed fraction $\hat{p}$?

$$\hat{p} = \frac{317}{400} \approx 0.79$$



2. Give error bars on this estimate.

We have $n = 400$ tosses of a coin of unknown bias p.

The distribution of $\hat{p}$ is $N\left(p, \frac{p(1-p)}{n}\right)$. (observed fraction of heads)

Standard deviation $= \sqrt{\frac{p(1-p)}{n}}$.

Is there a problem here?

We don't know p, so we can't evaluate $\sqrt{\frac{p(1-p)}{n}}$. Instead, use $\hat{p}$.

Since we don't know the true standard deviation $\sqrt{p(1-p)}$ of each sample, use the observed standard deviation $\sqrt{\hat{p}(1-\hat{p})}$.

$\hat{p} = 0.79$

- Estimate the standard deviation of $\hat{p}$.

$$\text{std}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.79 \times 0.21}{400}} \approx 0.02$$

- The normal approximation gives confidence intervals:
  - 68.3% interval: 0.79 ± 0.02 ← 1 standard dev
  - 95.5% interval: 0.79 ± 0.04 ← 2 standard devs
  - 99.7% interval: 0.79 ± 0.06 ← 3 deviations

  What is a 95% confidence interval for p?

  $[0.75, 0.83]$

- What does a "95% confidence interval" really mean?

95% is our confidence in the procedure, not in the specific interval it has output.

If we were to repeat the experiment over and over again, each time producing a confidence interval, then at least 95% of the time, these intervals would contain the true probability.

Estimating an average

400 numbers: 9, 13, 10, 8, . . .

In a certain town, a random sample is taken of 400 people age 25 and over. The average years of schooling of this sample is 11.6 years, with a standard deviation of 4.1. Find a 95% confidence interval for the average educational level of people 25 and over in this town.

Estimating an average

In a certain town, a random sample is taken of 400 people age 25 and over. The average years of schooling of this sample is 11.6 years, with a standard deviation of 4.1. Find a 95% confidence interval for the average educational level of people 25 and over in this town.

What is the distribution of the observed average?

- Let the true mean educational level (# years of schooling) be μ, with stddev σ.
- We draw n samples from this distribution, and take the average $\hat{\mu}$. $\hat{\mu} = 11.6$, $\hat{\sigma} = 4.1$

$n = 400$

Let the samples be $X_1, X_2, \ldots, X_n$ $(n = 400)$

X_1: # years of schooling of person 1; X_2: person 2; ...

unknown params: $\mathbb{E}[X_i] = \mu$, $\text{var}(X_i) = \sigma^2$

$\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$. By CLT, $\hat{\mu} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

$\therefore \text{std}(\hat{\mu}) = \dfrac{\sigma}{\sqrt{n}} \approx \dfrac{\hat{\sigma}}{\sqrt{n}}$ ← we don't know σ, so use $\hat{\sigma}$ instead

σ ← about 0.2

$$\text{std}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}} \approx \frac{\hat{\sigma}}{\sqrt{n}} = \frac{4.1}{\sqrt{400}} \approx 0.2$$

Estimated value $\hat{\mu} = 11.6$

95% confidence interval =

$$[\hat{\mu} - 2\,\text{std}(\hat{\mu}),\ \hat{\mu} + 2\,\text{std}(\hat{\mu})]$$

$$= [11.6 - 0.4,\ 11.6 + 0.4] = [11.2, 12.0]$$

Worksheet 12 # 1, 3, 5, 7, 8, 9, 10

3 (a) X_i = number of darts in wedge i

[Throw 100 darts, and there are 20 equal wedges]

Want $\mathbb{E}[X_i]$ and $\text{var}(X_i)$.

Toss a coin of bias $p = \frac{1}{20}$ (whether a particular dart falls in that wedge)

Toss the coin $n=100$ times (since 100 darts)

$$\mathbb{E}[X_i] = np = 100 \cdot \frac{1}{20} = 5$$

$$\text{var}(X_i) = np(1-p) = 100 \cdot \frac{1}{20} \cdot \frac{19}{20} = \frac{19}{4} = 4.75$$

(b) X_i is approximated by what normal distribution?

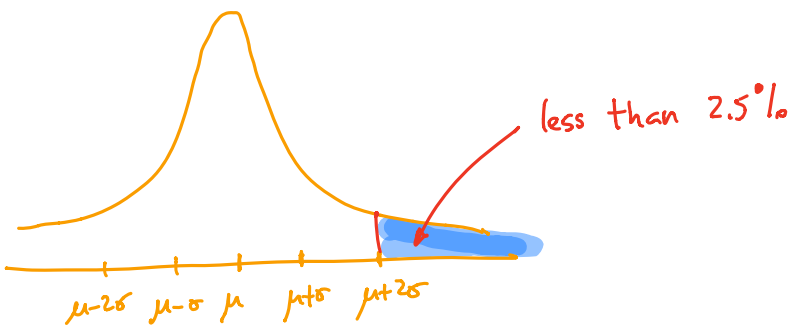
$N(5, 4.75)$

What is a 95% confidence interval for X?

$$5 \pm 2\sqrt{4.75} = 5 \pm 4.4$$

Just need an upper bound...

$$5 + 2\sqrt{4.75} = 9.4$$



(c) $Y_i = \begin{cases} +1 & \text{if } i^{th} \text{ dart is on red} \\ -1 & \text{black} \end{cases}$

$$\mathbb{E}[Y_i] = +1 \cdot \frac{1}{2} - 1 \cdot \frac{1}{2} = 0$$

$$\text{var}(Y_i) = 1$$

(d) Z_r = # darts on red

Z_b = # darts on black

Z_r = # darts on red

Z_b = # on black

$Z_r - Z_b = Y_1 + Y_2 + \cdots + Y_{100}$

$\left.\begin{array}{l} \mathbb{E}[Y_i] = 0 \\ \text{var}(Y_i) = 1 \end{array}\right\}$

What is a good normal approximation to $Z_r - Z_b$?

- $\mathbb{E}[Z_r - Z_b] = \mathbb{E}[Y_1 + \cdots + Y_{100}]$
  $= \mathbb{E}[Y_1] + \cdots + \mathbb{E}[Y_{100}] = 0$
- $\text{var}(Z_r - Z_b) = \text{var}(Y_1 + \cdots + Y_{100})$
  $= \text{var}(Y_1) + \cdots + \text{var}(Y_{100}) = 100$

Normal approximation: $Z_r - Z_b \sim N(0, 100)$.

(e) Give a 99% confidence interval for $Z = |Z_r - Z_b|$.

99% confidence interval for $Z_r - Z_b$ is $[-30, 30]$

99% confidence interval for Z is $[0, 30]$.