Sampling

DSE 210

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, and let X be the number of heads. What is $\mathbb{E}(X)$?

Review: Variance

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:

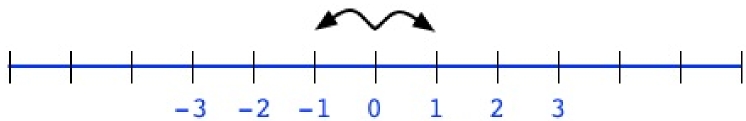
- $\text{var}(X_1 + \cdots + X_k) = \text{var}(X_1) + \cdots + \text{var}(X_k)$ if X_i's independent.
- $\text{var}(aX + b) = a^2\text{var}(X)$.

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.

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?

Variance of a sum

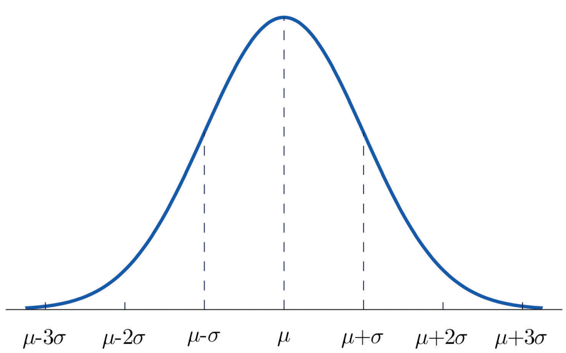
Symmetric random walk. A drunken man sets out from a bar. At each time step, he either moves one step to the right or one step to the left, with equal probabilities. Roughly where is he after n steps?



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 $\text{var}(X_i)$?
- What are $\mathbb{E}(X)$ and $\text{var}(X)$?
- What is $\text{std}(X)$?
- What is the distribution over his possible positions?

The normal 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

Suppose $X_1, \ldots, X_n$ are independent, each with mean μ and variance σ^2.

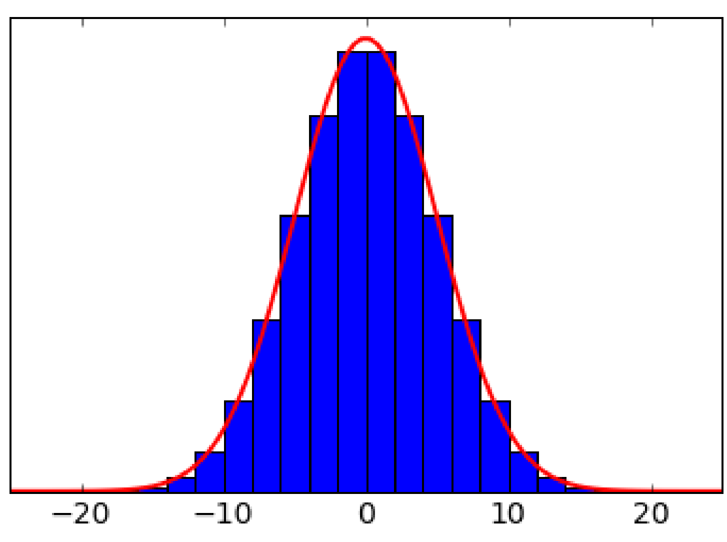
Let $S_n = X_1 + \cdots + X_n$. What are the mean and variance of S_n?

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** $(X_1 + \cdots + X_n)/n$?

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

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?
  Answer: $N(p, p(1-p)/n)$.
  Mean p, standard deviation on the order of $1/\sqrt{n}$.

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?

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

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?

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 $\widehat{p}$?

2. Give error bars on this estimate.

Is there a problem here?

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

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

- The normal approximation gives confidence intervals:
  - 68.3% interval: 0.79 ± 0.02
  - 95.5% interval: 0.79 ± 0.04
  - 99.7% interval: 0.79 ± 0.06

  What is a 95% confidence interval for p?

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

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 be μ, with stddev σ.
- We draw n samples from this distribution, and take the average $\widehat{\mu}$.