Statistics from samples and Limit theorems

Subsection 1

Statistics from *iid* samples

Where have we seen *iid* samples?

- Bernoulli trials
- Monte carlo simulations
- Computing histograms

Bernoulli trials

- Experiment and an event of interest A
  - Occurance of an Event A is considered success
  - What is $p = P(A)$?
- Bernoulli trails
  - Independent repetitions or trials of experiment, say, n times
  - $X_i = 1$ if A occurs in the i-th trial, and $X_i = 0$ otherwise

iid Bernoulli samples: $X_1, X_2, \ldots, X_n$

Goal: Try to estimate $P(A) = P(X_i = 1)$

- Useful in finding prevalence of a disease in a population etc.

Monte carlo simulations

- Experiment and event of interest A
  - Too complex for modeling and computation
  - Can be simulated on a computer
- Repeat simulation n times independently
  - Record number of occurances of A as n_A

$$P(A) \approx \frac{n_A}{n}$$

  - Why is the above true? What should be n?

iid samples: $X_1, X_2, \ldots, X_n$

- $X_i = 1$ if A occurs in the i-th trial, and $X_i = 0$ otherwise
- Estimate $P(A) = P(X_i = 1)$ (similar to Bernoulli trial)

Computing histograms

- n data points of some variable of interest
  - $x_1, x_2, \ldots, x_n$
- Bin: $[a, b]$
  - n_b: number of x_i that fall inside $[a, b]$

Model

- X: continuous random variable with density $f_X(x)$
- Event $A = (a < X < b)$

Histogram count

- Data points: *iid* samples $X_1, X_2, \ldots, X_n \sim f_X(x)$
- Estimate $P(a < X < b) \approx n_b/n$

iid samples hold information on distribution

- What is common to all 3 of the previous scenarios?
  - Given: *iid* samples
  - Goal: get some partial information about the distribution
    - ⋆ Procedures to gather the information needed
- *iid* samples of an unknown or partially known distribution form the input for statistical procedures
  - Data: modelled as observations from *iid* repetitions of an experiment
  - Example: Iris data
    - ⋆ Data from every iris is considered to be *iid* observations from the distribution of the 4 lengths

Analysis

- How to decide if the statistical procedure is "good"?
- How many samples are needed for a "goodness" guarantee?

Example: 20 *iid* Bernoulli(p) samples with p unknown

- **Goal**
  - Find p from *iid* samples
- Sampling 1
  - 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1
- Sampling 2
  - 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1
- Sampling 3
  - 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
- and so on. . . .
- **Important**
  - p is the same for all samplings
  - However, samples do not remain the same
  - Each sample in each sampling is an observation of a random variable
- **Requirement on statistical procedure**
  - In spite of variations in samples, provide p with some guarantee

What is a typical statistical problem?

- **Model for Samples**: $X_1, X_2, \ldots, X_n \sim$ iid X
- **Given "data"**: $x_1, x_2, \ldots, x_n$ from one sampling instance
- Distribution of X is partially known or unknown
  - What is partially known? Know distribution but parameters unknown
  - Example: Bernoulli(p) with p unknown, Normal(μ, σ^2) with μ and σ unknown
- **Goal**: Procedures to find information about the distribution of X
- **What information?**
  - What is the mean of X? What is the variance of X?
  - What is $P(X > t)$? What is $P(a < X < b)$?
  - What is the distribution of X? What is the size of T_X?

Subsection 2

Empirical distribution and descriptive statistics

Empirical distribution

Definition (Empirical distribution)

Let $X_1, X_2, \ldots, X_n \sim X$ be iid samples. Let $\#(X_i = t)$ denote the number of times t occurs in the samples. The empirical distribution is the discrete distribution with PMF

$$p(t) = \frac{\#(X_i = t)}{n}.$$

Example: $n = 20$

- 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1
  - $p(0) = 8/20$, $p(1) = 12/20$
- 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1
  - $p(0) = 7/20$, $p(1) = 13/20$
- 1, 2, 0, 3, 0, 0, 1, 2, 0, 1, 3, 2, 1, 1, 0, 3, 0, 2, 2, 1
  - $p(0) = 6/20$, $p(1) = 6/20$, $p(2) = 5/20$, $p(3) = 3/20$

$6/20\ 6/20\ 5/20\ 3/20$
$\{0, 1, 2, 3\}$

Observations about empirical distribution

- Is the empirical distribution random?
  - Yes, it depends on the actual sample instances
  - t and $p(t)$ may change from one sampling to another
  - Example: 20 Bernoulli(p) samples
- **Descriptive statistics**: Properties of empirical distribution
  - Mean of the distribution
  - Variance of the distribution
  - Probability of an event
- As number of samples increases, the properties of empirical distribution *should* become close to that of the original distribution

Sample mean

Definition (Sample mean)

Let $X_1, X_2, \ldots, X_n$ be iid samples. The sample mean, denoted $\overline{X}$, is defined to be the random variable

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

- Given a sampling $x_1, \ldots, x_n$, the value taken by the sample mean $\overline{X}$ is $\overline{x} = (x_1 + \cdots + x_n)/n$. Often, $\overline{X}$ and $\overline{x}$ are both called sample mean.

Example: $n = 20$

- 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1
  - Value taken by $\overline{X}$: 12/20
- 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1
  - Value taken by $\overline{X}$: 13/20
- 1, 2, 0, 3, 0, 0, 1, 2, 0, 1, 3, 2, 1, 1, 0, 3, 0, 2, 2, 1
  - Sample mean: 25/20

Illustration 1: Bernoulli(0.5) samples

$$X_1, \ldots, X_n \sim \text{ iid } \{ \overset{1/2}{0} , \overset{1/2}{1} \}$$

Distribution mean = 1/2

- $n = 5$
  - Samples: 0, 0, 1, 1, 1; Sample mean: 3/5
  - Samples: 1, 1, 1, 0, 1; Sample mean: 4/5
  - Samples: 0, 1, 1, 1, 0; Sample mean: 3/5
- $n = 20$
  - 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1; 12/20
  - 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1; 13/20
  - 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1; 13/20
- $n = 200$
  - Sampling 1: 95/200, Sampling 2: 102/200, Sampling 3: 98/200
- $n = 1000$
  - Sampling 1: 495/1000, Sampling 2: 490/1000, Sampling 3: 504/1000

Illustration 2: Normal(0, 1) samples

$$X_1, \ldots, X_n \sim \text{ iid Normal}(0, 1)$$

0 ↑ Distribution mean

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

- $n = 5$
  - Samples: 2.17, 0.10, -0.75, -1.05 , -1.72; Sample mean: -0.25
  - Samples: -0.26, 0.12, -0.31, -0.07, 1.35; Sample mean: 0.17
  - Samples: -0.20, 0.37, 1.00, -0.41, -0.21; Sample mean: 0.11
- $n = 20$
  - Sampling 1: 0.08, Sampling 2: -0.24, Sampling 3: 0.41
- $n = 200$
  - Sampling 1: -0.01, Sampling 2: 0.11, Sampling 3: -0.12
- $n = 1000$
  - Sampling 1: 0.04, Sampling 2: -0.04, Sampling 3: -0.02

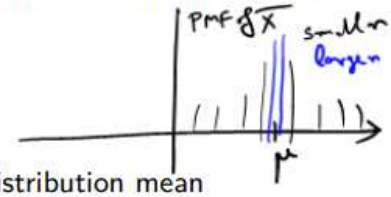
Expected value and variance of sample mean

Theorem

Let $X_1, X_2, \ldots, X_n$ be iid samples whose distribution has a finite mean μ and variance σ^2. The sample mean $\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ has expected value and variance given by

$$E[\overline{X}] = \mu, Var(\overline{X}) = \frac{\sigma^2}{n}.$$

- Expected value of sample mean equals the expected value or mean of the distribution
  - Mean of distribution: constant real number and not random
  - Sample mean: random variable with mean equal to distribution mean
- Variance of sample mean decreases with n
  - As n increases...
    - ★ variance of sample mean tends to zero
    - ★ the spread of sample mean will decrease
    - ★ sample mean will take values close to the distribution mean



Sample variance

Definition (Sample variance)

Let $X_1, X_2, \ldots, X_n$ be iid samples. The sample variance, denoted S^2, is defined to be the random variable

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1},$$

where $\overline{X}$ is the sample mean.

- Given a sampling $x_1, \ldots, x_n$, the value taken by the sample variance S^2 is $s^2 = ((x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2)/(n-1)$. Often, S^2 and s^2 are both called sample variance.
- Why $n-1$ in the denominator instead of n?
  - Some books use n (this causes confusion)
  - Expected value of sample variance is simple in this case

Expected value of sample variance

Theorem

Let $X_1, X_2, \ldots, X_n$ be iid samples whose distribution has a finite variance σ^2. The sample variance $S^2 = \frac{(X_1-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n}$ has expected value given by

$$E[S^2] = \sigma^2.$$

- Expected value of sample variance equals the variance of the distribution
  - Variance of distribution: constant real number and not random
  - Sample variance: random variable with mean equal to distribution variance
- Values of sample variance, on average, give the variance of distribution
  - Variance of sample variance will decrease with number of samples (in most cases)
  - As n increases, sample variance takes values close to distribution variance

Illustration

- Bernoulli(1/2), mean $= 0.5$, variance $= 0.25$
  - Sample variance values: $n = 20$
    - 0.26, 0.26, 0.26, 0.25, 0.26
  - Sample variance values: $n = 200$
    - 0.2500, 0.2487, 0.2496, 0.2456, 0.2476
  - Sample variance values: $n = 1000$
    - 0.2498, 0.2490, 0.2499, 0.2501, 0.2502
- Normal(0,1), mean $= 0$, variance $= 1$
  - Sample variance values: $n = 20$
    - 0.89, 0.57, 1.19, 1.01, 1.41
  - Sample variance values: $n = 200$
    - 0.93, 1.07, 0.85, 0.83, 1.09
  - Sample variance values: $n = 1000$
    - 1.0268, 0.9535, 0.9781, 0.9766, 0.9831

Sample proportion

$$X_1, X_2, \ldots, X_n \sim X$$

- *iid* samples from the distribution of X
- Let A be an event defined using X
  - Example: $A = (X > t)$, $A = (a < X < b)$ etc

Definition (Sample proportion)

The sample proportion of A, denoted $S(A)$, is defined as

$$S(A) = \frac{\#(X_i \text{ for which } A \text{ is true})}{n}.$$

- Samples: 0, 1, 1, 1, 0
  - $S(X = 1) = 3/5$
- Samples: -0.2, 1.1, 0.3, -1.2, 0.7
  - $S(X \leq 0) = 2/5$, $S(0 < X < 1) = 2/5$, $S(X > 1) = 1/5$

Expected value and variance of sample proportion

Theorem

Let $X_1, X_2, \ldots, X_n$ be iid samples from the distribution of X. Let A be an event defined using X and let $P(A)$ be the probability of A. The sample proportion of A, denoted $S(A)$, has expected value and variance given by

$$E[S(A)] = P(A), Var(S(A)) = \frac{P(A)(1 - P(A))}{n}.$$

Proof

- Convert samples into Bernoulli($P(A)$) samples $Y_1, \ldots, Y_n$
  - $Y_i = 1$ if A is true for X_i, and $Y_i = 0$ otherwise
- $S(A)$ is the sample mean of $Y_1, \ldots, Y_n$
- As n increases, values of $S(A)$ will be close to $P(A)$
  - Mean of $S(A)$ equals $P(A)$
  - Variance of $S(A)$ tends to 0

Illustration

$X_1, \ldots, X_n \sim \text{Normal}(0, 1)$

- $P(X \leq -1) = 0.159$ ← from distribution $\int_{-\infty}^{-1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.159\ldots$
  - Sample proportion values: $n = 20$
    - ⋆ 0.15, 0.20, 0.15, 0.15, 0.15 (5 samplings)
  - Sample proportion values: $n = 200$
    - ⋆ 0.170, 0.140, 0.150, 0.155, 0.165 (5 samplings)
  - Sample proportion values: $n = 1000$
    - ⋆ 0.160, 0.180, 0.162, 0.135, 0.153 (5 samplings)
- $P(-1 < X < 1) = 0.683$
  - Sample proportion values: $n = 20$
    - ⋆ 0.75, 0.70, 0.55, 0.45, 0.70
  - Sample proportion values: $n = 200$
    - ⋆ 0.705, 0.690, 0.705, 0.670, 0.720
  - Sample proportion values: $n = 1000$
    - ⋆ 0.678, 0.678, 0.686, 0.679, 0.681

Where have we seen *iid* samples?

- Bernoulli trials: Sample mean tends to distribution mean
  - Bernoulli(p) samples
  - Distribution mean $= p$
  - Sample mean = fraction of successes
- Monte carlo simulations
  - Sample proportion tends to actual probability
- Computing histograms
  - Sample proportion tends to actual probability

Subsection 3

Illustrations with data

Iris data

- 3 classes of irises: 0, 1, 2
  - 50 instances of data for each class
  - Each instance: [sepal length, sepal width, petal length, petal width] (cm)
- Sepal length of Class 0
  - Model: *iid* samples according to some unknown distribution
  - Data: 5.1, 4.9, 4.7, . . . , 5.3, 5
  - Sample mean: 5.006, Sample variance: $0.1242 = 0.3524^2$
  - S(Sepal length > 5) $= 22/50$, S($4.8 <$ Sepal length < 5.2) $= 20/50$
- Petal width of Class 3
  - Model: *iid* samples according to some unknown distribution
  - Data: 2.5, 1.9, 2.1, . . . , 2.3, 1.8
  - Sample mean: 2.026, Sample variance: $0.0754 = 0.2746^2$
  - S(Petal width > 2) $= 23/50$, S($1.8 <$ Petal length < 2.2) $= 17/50$
- Model: how good is the iid samples model?

Taj Mahal air quality

Date	SO2	NO2	PM2.5	PM10
12/4	4	60	77	185
13/4	4	53	65	196
11/4	4	57	72	223
10/4	4	45	68	200
8/4	5	33	52	250
7/4	4	27	67	266
6/4	4	12	60	219
5/4	7	27	70	207
4/4	4	58	100	282
3/4	4	17	55	158
1/4	4	31	37	465
Max	80	80	60	100

- 24-hour average of particles in air, units: micrograms/cubic metre

Taj Mahal air quality sample statistics

- Sample means
  - SO2: 4.36, NO2: 38.18, PM2.5: 65.72, PM10: 241
- Sample standard deviations
  - SO2: 0.9244^2, NO2: 17.1803^2, PM2.5: 15.9002^2, PM10: 82.6184^2
- S(max exceeded)
  - SO2: 0, NO2: 0, PM2.5: 7/11, PM10: 11/11
- Model
  - Do you like the iid samples model for this data?
  - Is the Taj in trouble or not? How do we answer such questions?
  - How confident are our conclusions when we have looked at just 11 data points?

IPL: Runs scored in Deliveries 0.1, 0.2, 0.3

- Data from 1598 innings
  - See shared spreadsheet
  - Download csv from cricsheet.org
- All calculations done using spreadsheets or other computing tools
- Sample means
  - 0.1: 0.7347, 0.2: 0.8686, 0.3: 0.9524
- Sample variances
  - 0.1: 1.4975, 0.2: 1.7961, 0.3: 2.0666
- Sample proportions
  - S(dot ball) - 0.1: 0.5989, 0.2: 0.5551, 0.3: 0.5338
  - S(4 or 6) - 0.1: 0.0914, 0.2: 0.1145, 0.3: 0.1302
- Clear trend from samples
  - Runs scored increases from 0.1 to 0.3
- Enough data points to be confident in the trend
  - Agrees with intuition
- Model: Do you like the iid samples model for each delivery?

Subsection 4

Sum of independent random variables

Expected value and variance

Theorem

Let $X_1, X_2, \ldots, X_n$ be random variables. Let $S = X_1 + \cdots + X_n$ be their sum. Then,

$$E[S] = E[X_1] + \cdots + E[X_n].$$

If $X_1, \ldots, X_n$ are pairwise uncorrelated, then

$$Var(S) = Var(X_1) + \cdots + Var(X_n).$$

- What is pairwise uncorrelated? $E[X_i X_j] = E[X_i]E[X_j]$ for all i, j, $i \neq j$
- Mean of sum is sum of means
- If uncorrelated, variance of sum is sum of variances
- If the X_i are independent, they are also uncorrelated
  - So, above result holds for independent random variables

Extensions of previous result

- Scaling and summing
  - Suppose $S = a_1 X_1 + \cdots + a_n X_n$, where a_i are constants
  - $E[S] = a_1 E[X_1] + \cdots + a_n E[X_n]$
  - $\text{Var}(S) = a_1^2 \text{Var}(X_1) + \cdots + a_n^2 \text{Var}(X_n)$, if uncorrelated
- *iid* samples: $X_1, \ldots, X_n \sim X$, iid
  - Suppose $S = a_1 X_1 + \cdots + a_n X_n$, where a_i are constants
  - $E[S] = (a_1 + \cdots + a_n)E[X]$
  - $\text{Var}(S) = (a_1^2 + \cdots + a_n^2)\text{Var}(X)$
- Sample mean: $X_1, \ldots, X_n \sim X$, iid
  - $\overline{X} = (X_1 + \cdots + X_n)/n$, $a_i = 1/n$
  - $E[\overline{X}] = E[X]$
  - $\text{Var}(\overline{X}) = \text{Var}(X)/n$

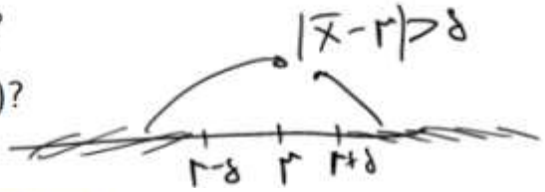
Sample mean versus distribution mean

$$X_1, \ldots, X_n \sim \text{ iid } X$$

- Let $\mu = E[X]$, $\sigma^2 = \text{Var}(X)$
- Sample mean: $\overline{X} = (X_1 + \cdots + X_n)/n$
  - Expected value: μ, Variance: σ^2/n
  - Variance (or spread) goes to 0 as n grows

Can we say something more precise about $\overline{X}$ and μ?

- What is $P(\overline{X} > \mu + \delta)$?
- What is $P(\overline{X} < \mu - \delta)$?
- What is $P(|\overline{X} - \mu| > \delta)$?



Weak law of large numbers

$$X_1, \ldots, X_n \sim \text{ iid } X$$

$\overline{X}$: converges to μ "in probability"

- Let $\mu = E[X]$, $\sigma^2 = \text{Var}(X)$
- Sample mean: $\overline{X} = (X_1 + \cdots + X_n)/n$
  - Expected value: μ, Variance: σ^2/n

Theorem (Weak law of large numbers)

$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2} \to 0.$$

- With probability more than $1 - \frac{\sigma^2}{n\delta^2}$, sample mean lies in $[\mu - \delta, \mu + \delta]$
  - **What is the meaning of this probability?**
- Chebyshev is usually a very "weak" bound, and we will see sharper bounds soon

Examples: n iid samples

- Bernoulli(p) samples
  - With probability more than $1 - \frac{p(1-p)}{n\delta^2}$, sample mean lies in $[p - \delta, p + \delta]$
- Uniform$\{-M, \ldots, M\}$ samples
  - With probability more than $1 - \frac{M(M+1)}{3n\delta^2}$, sample mean lies in $[-\delta, \delta]$
- Normal$(0, \sigma^2)$ samples
  - With probability more than $1 - \frac{\sigma^2}{n\delta^2}$, sample mean lies in $[-\delta, \delta]$
- Uniform$[-A, A]$ samples
  - With probability more than $1 - \frac{A^2}{3n\delta^2}$, sample mean lies in $[-\delta, \delta]$
- When distribution is known, a precise statement is possible about "confidence" of finding sample mean within a certain precise interval
  - Improvement in bound will improve precision

Examples: Iris, Taj Mahal and IPL

- Iris data: Sepal length
  - $n = 50$, Sample mean: 5.006, Sample variance: 0.1242
  - With probability more than $1 - \sigma^2/50\delta^2$, sample mean lies in $[\mu - \delta, \mu + \delta]$
    - Works for $\delta > \sigma/\sqrt{50}$
- Taj Mahal air quality: PM2.5
  - $n = 11$, Sample mean $= 65.72$, Sample variance $= 15.9^2$
  - With probability more than $1 - \sigma^2/11\delta^2$, sample mean lies in $[\mu - \delta, \mu + \delta]$
    - Works for $\delta > \sigma/\sqrt{11}$
- IPL: Runs scored in Delivery 0.3
  - $n = 1598$, Sample mean: 0.9524, Sample variance: 2.0666
  - With probability more than $1 - \sigma^2/1598\delta^2$, sample mean lies in $[\mu - \delta, \mu + \delta]$
    - Works for $\delta > \sigma/\sqrt{1598}$
- What to do when distribution is unknown? Have to assume something

Subsection 5

Sum of independent random variables II

Subsection 6

Concentration phenomenon
Bounding $P(|\overline{X} - \mu| > t)$

$$X_1, \ldots, X_n \sim \text{iid } X$$

- Sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$
- Chebyshev bound

$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2}$$

How weak is Chebyshev?

- Let $X \sim$ Bernoulli$(1/2)$, $\mu = 0.5$, $\sigma^2 = 0.25$
  - $n = 10$: $P(|\overline{X} - 0.5| > 0.3) = 0.0215 \leq 0.278$
  - $n = 50$: $P(|\overline{X} - 0.5| > 0.3) = 5.61 \times 10^{-6} \leq 0.056$
- Chebyshev falls as $1/n$
- In many cases, we can have e^{-cn}
  - Exponential fall with n
  - Much much faster than $1/n$

Markov, Chebyshev and Chernoff

- Markov inequality: X takes positive values

$$P(X > t) \leq \frac{E[X]}{t}$$

- Chebyshev inequality: X could take positive/negative values with finite variance

$$P(|X - E[X]| > t) = P((X - E[X])^2 > t^2) \leq \frac{\text{Var}(X)}{t^2}$$

- Chernoff inequality: X could take positive/negative values, $E[X] = 0$

$$P(X > t) = P(e^{\lambda X} > e^{\lambda t}) \leq \frac{E[e^{\lambda X}]}{e^{\lambda t}}, \ \lambda > 0$$

- **Moment generating function (MGF)** of X: $E[e^{\lambda X}]$ (for $E[X] = 0$)
  - Pick λ that provides best bound
  - MGF too unwieldy? Use upper bound on MGF

MGF of Centralised Bernoulli(1/2)

- What is centralising?
  - X: random variable with mean $E[X]$
  - Centralised version of X: $X - E[X]$
- Centralised Bernoulli(1/2)
  - Bernoulli(1/2): $\{\overset{1/2}{0}, \overset{1/2}{1}\}$, mean $= 1/2$
  - Centralised: $X \sim \{\overset{1/2}{-1/2}, \overset{1/2}{1/2}\}$, mean $= 0$

$$E[e^{\lambda X}] = P(X=-1/2)e^{-\lambda/2} + P(X=1/2)e^{\lambda/2} = \frac{e^{\lambda/2} + e^{-\lambda/2}}{2}$$

Bound on MGF

$$E[e^{\lambda X}] = \frac{e^{\lambda/2} + e^{-\lambda/2}}{2} \leq e^{\lambda^2/4}$$

MGF and sum of *iid* random variables

$$X_1, \ldots, X_n \sim \text{iid } X$$

- Let $S = X_1 + \cdots + X_n$. What is MGF of S?
- Why is this question important?
  - MGF gives bounds on $P(S > t)$
  - Sample mean $\overline{X} = S/n$

$$E[e^{\lambda S}] = E[e^{\lambda X_1} \cdots e^{\lambda X_n}] = E[e^{\lambda X_1}] \cdots E[e^{\lambda X_n}] = E[e^{\lambda X}]^n$$

- **MGF of sum of independent random variables is product of the individual MGFs**

Example: Sum of centralised Benoulli

- $X \sim$ Centralised Bernoulli(1/2), i.e. $\{-1/2, 1/2\}^{1/2}$

$$X_1, \ldots, X_n \sim \text{iid } X$$

- $S = X_1 + \cdots + X_n$

$$E[e^{\lambda S}] = \left(\frac{e^{\lambda/2} + e^{-\lambda/2}}{2}\right)^n \leq e^{n\lambda^2/4}$$

- **Upper bound is so much easier than MGF!**

Chernoff bound for Binomial(n,1/2)

$$X_1, \ldots, X_n \sim \text{iid Centralised Bernoulli}(1/2)$$

- $S = X_1 + \cdots + X_n$

$$P(S > t) \leq \frac{E[e^{\lambda S}]}{e^{\lambda t}} \leq e^{n\lambda^2/4 - \lambda t}$$

- Pick $\lambda = 2t/n$: $P(S > t) \leq e^{-t^2/n}$
- Now, $Y_1 = X_1 + 1/2 \sim$ Bernoulli(1/2). So, $Y = Y_1 + \cdots + Y_n = S + n/2 \sim$ Binomial(n,1/2)

$$P(Y > n/2 + n\delta/2) = P(S > n\delta/2) \leq e^{-n\delta^2/4}$$

- Chebyshev: $P(Y > n/2 + n\delta/2) \leq \frac{1}{n\delta^2}$

Chebyshev vs Chernoff: $Y \sim \text{Binomial}(n, 1/2)$

n	Event, $\delta = 0.6$	Prob	$1/n\delta^2$	$e^{-n\delta^2/4}$
10	$Y - 5 > 5 \times 0.6$	0.0107	0.278	0.407
50	$Y - 25 > 25 \times 0.6$	2.81×10^{-6}	0.056	0.011
100	$Y - 50 > 50 \times 0.6$	1.35×10^{-10}	0.028	1.23×10^{-4}
200	$Y - 100 > 100 \times 0.6$	4.16×10^{-19}	0.014	1.52×10^{-8}
400	$Y - 200 > 200 \times 0.6$	5.40×10^{-36}	0.007	2.32×10^{-16}

- $1/n$ vs e^{-cn}: difference is clearly seen as n increases
  - $1/n$ is giving a very wrong idea about the magnitude of the probability

Remarks on concentration phenomenon

$X_1, \ldots, X_n \sim$ iid X, $Y = X_1 + \cdots + X_n$

- Concentration phenomenon
  - Exponential bounds for $P(Y > E[Y] + t)$ by bounding $E[e^{\lambda Y}]$ and using Chernoff
  - We saw this when X is Bernoulli
- What about $P(Y < E[Y] - t)$?
  - Replace Y with $-Y$ on $P(Y > E[Y] + t)$
- What about other distributions? Several extensions exist
  - Hoeffding's inequality: X is bounded within an interval $[-M, M]$
  - Bennett's inequality: X is bounded in $[-M, M]$ and has finite variance
- Y is the sum of the iid samples. What about other functions $f(X_1, \ldots, X_n)$?
  - Many extensions: f should depend "equally" on all variables

Subsection 7

Central Limit Theorem

Moment generating function (MGF)

Definition (MGF)

Let X be a zero-mean random variable. The MGF of X, denoted $M_X(\lambda)$, is a function from R to R defined as

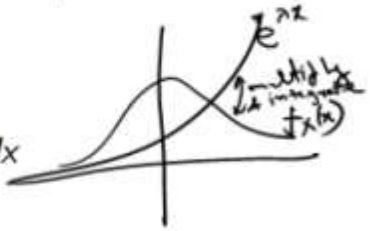
$$M_X(\lambda) = E[e^{\lambda X}].$$

- X: Discrete with PMF f_X
  - X takes values $\{x_1, x_2, \ldots\}$

$$M_X(\lambda) = f_X(x_1)e^{\lambda x_1} + f_X(x_2)e^{\lambda x_2} + \cdots$$

- X: continuous with PDF f_X and support T_X

$$M_X(\lambda) = \int_{x \in T_X} f_X(x)e^{\lambda x} dx$$



Examples

- $X \sim \{0\}$ (X is 0 with probability 1)
  - $M_X(\lambda) = 1 \times e^0 = 1$
- $X \sim \{-p, 1-p\}$
  - $M_X(\lambda) = (1-p)e^{-p\lambda} + pe^{(1-p)\lambda}$
- $X \in \{-1, 0, 2\}$
  - $M_X(\lambda) = 0.5e^{-\lambda} + 0.25 + 0.25e^{2\lambda}$
- $M_X(\lambda) = (1/3)e^{3\lambda/2} + (1/6)e^{-3\lambda} + (1/8)e^{-\lambda} + (1/8)e^{\lambda} + 1/4$
  - $X \sim \{-3, -1, 0, 1, 3/2\}$
- $X \sim \text{Normal}(0, \sigma^2)$
  - $M_X(\lambda) = e^{\lambda^2\sigma^2/2}$

Why Moment Generating Function?

$$E[e^{\lambda X}] = E[1 + \lambda X + \frac{\lambda^2}{2!}X^2 + \frac{\lambda^3}{3!}X^3 + \cdots]$$
$$= 1 + \lambda E[X] + \frac{\lambda^2}{2!}E[X^2] + \frac{\lambda^3}{3!}E[X^3] + \cdots$$

- $X \sim \text{Normal}(0, \sigma^2)$, $M_X(\lambda) = e^{\lambda^2\sigma^2/2}$

$$1 + \lambda E[X] + \frac{\lambda^2}{2!}E[X^2] + \frac{\lambda^3}{3!}E[X^3] + \cdots = 1 + \frac{\lambda^2}{2!}\sigma^2 + \frac{\lambda^4}{4!}(3\sigma^4) + \cdots$$

- $E[X] = 0$, $E[X^2] = \sigma^2$, $E[X^3] = 0$, $E[X^4] = 3\sigma^4$ and so on

Examples: Sum of two independent random variables

$$X_1, X_2 \sim \text{iid } X, Y = X_1 + X_2$$

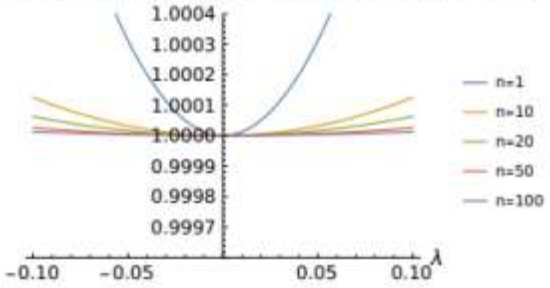
- $X \sim$ iid Bernoulli(p)
  - $M_X(\lambda) = (1-p)e^{-p\lambda} + pe^{(1-p)\lambda}$
  - $M_Y(\lambda) = M_X(\lambda)^2 = (1-p)^2e^{-2p\lambda} + 2p(1-p)e^{(1-2p)\lambda} + p^2e^{2(1-p)\lambda}$
    - $Y \sim \{-2p, 1-2p, 2(1-p)\}$ with probabilities $(1-p)^2$, $2p(1-p)$, p^2
- $X \in \{-1, 0, 2\}$ with probabilities $1/2$, $1/4$, $1/4$
  - $M_X(\lambda) = 0.5e^{-\lambda} + 0.25 + 0.25e^{2\lambda}$
  - $M_Y(\lambda) = 0.25e^{-2\lambda} + 0.25e^{-\lambda} + 0.0625 + 0.25e^{\lambda} + 0.125e^{2\lambda} + 0.0625e^{4\lambda}$
- $M_X(\lambda) = (1/3)e^{3\lambda/2} + (1/6)e^{-3\lambda} + (1/8)e^{-\lambda} + (1/8)e^{\lambda} + 1/4$
  - $M_Y(\lambda) = \frac{e^{-6t}}{36} + \frac{e^{-4t}}{24} + \frac{e^{-3t}}{12} + \frac{11e^{-2t}}{192} + \frac{1}{9}e^{-3t/2} + \frac{e^{-t}}{16} + \frac{3}{32} + \frac{e^{t/2}}{12} + \frac{e^{t}}{16} + \frac{1}{6}e^{3t/2} + \frac{e^{2t}}{64} + \frac{1}{12}e^{5t/2} + \frac{e^{3t}}{9}$

Example: MGF of sample mean

- Samples: $X_1, \ldots, X_n \sim$ iid X, $M_X(\lambda) = \frac{e^{\lambda/2}+e^{-\lambda/2}}{2}$
  - Sample mean: $\overline{X} = (X_1 + \cdots + X_n)/n$
  - $M_{X/n}(\lambda) = \frac{e^{\lambda/2n}+e^{-\lambda/2n}}{2}$

$$M_{\overline{X}}(\lambda) = \left(\frac{e^{\frac{\lambda}{2n}} + e^{-\frac{\lambda}{2n}}}{2}\right)^n$$

MGF of sample mean for different n



$M_{\overline{X}}(\lambda) \to 1$ as n increases

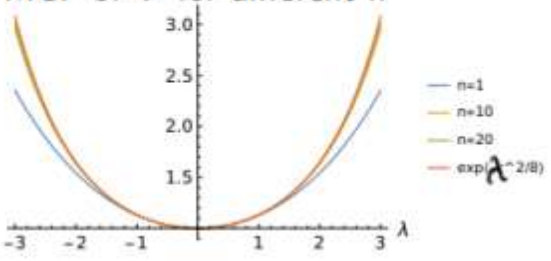
- WLLN: $\overline{X} \to E[X] = 0$
- Constant 0 has MGF $= 1$

MGF convergence at $1/\sqrt{n}$ scaling

- Samples: $X_1, \ldots, X_n \sim$ iid X, $M_X(\lambda) = \frac{e^{\lambda/2}+e^{-\lambda/2}}{2}$
  - $E[X] = 0$, $\mathrm{Var}(X) = 1/4$ ~~$= \sigma^2$~~
  - Consider $Y = (X_1 + \cdots + X_n)/\sqrt{n}$
  - $M_{X/\sqrt{n}}(\lambda) = \frac{e^{\lambda/2\sqrt{n}}+e^{-\lambda/2\sqrt{n}}}{2}$

$$M_Y(\lambda) = \left(\frac{e^{\frac{\lambda}{2\sqrt{n}}} + e^{-\frac{\lambda}{2\sqrt{n}}}}{2}\right)^n$$

MGF of Y for different n



$M_{\overline{X}}(\lambda) \to e^{\lambda^2\sigma^2/2}$ as n increases

$\updownarrow$

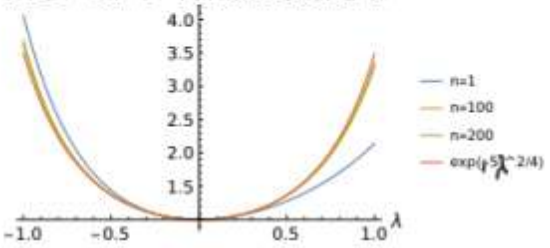
Normal $(0, \sigma^2)$

Another example: MGF convergence at $1/\sqrt{n}$ scaling

- Samples: $X_1, \ldots, X_n \sim$ iid X,
  $M_X(\lambda) = (1/3)e^{3\lambda/2} + (1/6)e^{-3\lambda} + (1/8)e^{-\lambda} + (1/8)e^{\lambda} + 1/4$
  - $E[X] = 0$, $\mathrm{Var}(X) = 5/2$ ~ σ^2
  - Consider $Y = (X_1 + \cdots + X_n)/\sqrt{n}$

$$M_Y(\lambda) = \left((1/3)e^{\frac{3\lambda}{2\sqrt{n}}} + (1/6)e^{-\frac{3\lambda}{\sqrt{n}}} + (1/8)e^{-\frac{\lambda}{\sqrt{n}}} + (1/8)e^{\frac{\lambda}{\sqrt{n}}} + 1/4\right)^n$$

MGF of Y for different n



$M_{\overline{X}}(\lambda) \to e^{\lambda^2\sigma^2/2}$ as n increases

$\updownarrow$

Normal$(0, \sigma^2)$

Central Limit Theorem (CLT)

Theorem (CLT)

Let $X_1, \ldots, X_n \sim$ iid X with $E[X] = 0$, $Var(X) = \sigma^2$. Let $Y = (X_1 + \cdots + X_n)/\sqrt{n}$. Then,

$$M_Y(\lambda) \to e^{\lambda^2\sigma^2/2}$$

- MGF of Normal$(0, \sigma^2)$: $e^{\lambda^2\sigma^2/2}$
- Y is said to converge in distribution to Normal$(0, \sigma^2)$

Observations

- Contrast with WLLN
  - $\overline{X} = (X_1 + \cdots + X_n)/n$ converges in distribution to $E[X]$.
- Sum of iid random variables tends to be normal
  - Do I like above statement? Not entirely.
  - Scaling is important: $1/n$ constant, $1/\sqrt{n}$ normal

Using CLT to approximate probability

$$X_1, \ldots, X_n \sim X$$

- Let $\mu = E[X]$, $\sigma^2 = \text{Var}(X)$
- $Y = X_1 + \cdots + X_n$
- What is $P(Y - n\mu > \delta n\mu)$?

Approximating using CLT

- $(Y - n\mu)/\sqrt{n}$: approximately Normal$(0, \sigma^2)$

$$\frac{Y - n\mu}{\sqrt{n}\sigma} \approx \text{Normal}(0, 1)$$

- $F(z)$: CDF of Normal(0,1) (known)
- $P(Y - n\mu > \delta n\mu) = P(\frac{Y - n\mu}{\sqrt{n}\sigma} > \frac{\delta\sqrt{n}\mu}{\sigma}) \approx 1 - F(\frac{\delta\sqrt{n}\mu}{\sigma})$

Approximating Binomial$(n, 1/2)$

$X_1, \ldots, X_n \sim$ iid Bernoulli(Y_2)

- $\mu = \sigma = 1/2$
- $P(Y - n/2 > 0.6n/2) \approx 1 - F(0.6\sqrt{n})$

F: CDF of Normal$(0,1)$

n	Event, $\delta = 0.6$	Prob	$1 - F(\sqrt{n}\delta)$
10	$Y - 5 > 5 \times 0.6$	0.0107	0.0289
50	$Y - 25 > 25 \times 0.6$	2.81×10^{-6}	1.10×10^{-5}
100	$Y - 50 > 50 \times 0.6$	1.35×10^{-10}	9.87×10^{-10}

- Approximation is really close!
- Normal approximation is quite good for Binomial

Examples

$$X_1, \ldots, X_n \sim \text{iid } X,\ Y = X_1 + \cdots + X_n$$

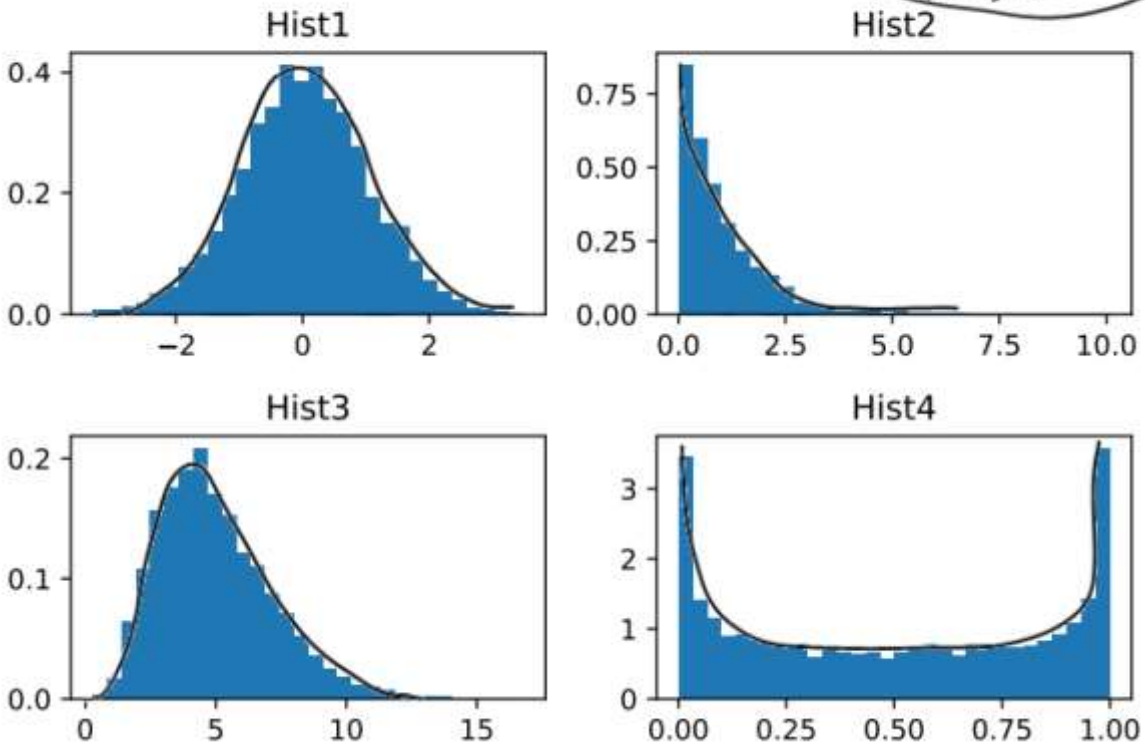
- $X \sim \{-3, -1, 0, 1, 3/2\}$ with probabilities $1/6, 1/8, 1/4, 1/8, 1/3$
  - $\mu = 0$, $\sigma^2 = 5/2$
  - CLT: $\frac{Y}{\sqrt{5n/2}} \approx \text{Normal}(0,1)$
  - $P(Y > \delta n) = P(\frac{Y}{\sqrt{5n/2}} > \delta\sqrt{2n/5}) \approx 1 - F(\delta\sqrt{2n/5})$
    - $n = 10$, $\delta = 1$: ≈ 0.0228
    - $n = 100$, $\delta = 1$: $\approx 1.27 \times 10^{-10}$
- $X \sim \text{Uniform}[-1, 1]$ (continuous)
  - $\mu = 0$, $\sigma^2 = 1/3$
  - CLT: $\sqrt{3}Y \approx \text{Normal}(0,1)$
  - $P(Y > 0.1\sqrt{n}) = P(\sqrt{3}Y > 0.1\sqrt{3n}) \approx 1 - F(0.1\sqrt{3n})$
    - $n = 10$: ≈ 0.2919
    - $n = 100$: ≈ 0.0416

Subsection 8

Distributions, properties and connections

Shapes of histograms: What distribution?

shape, location, rate, scale



Linear combination of iid normals

$$X_1, \ldots, X_n \sim \text{iid Normal}$$

- Let $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$
- Suppose $Y = a_1 X_1 + \cdots + a_n X_n$
  - Linear combination of iid normals
- Then,

$$Y \sim \text{Normal}(\mu, \sigma^2)$$

where $\mu = E[Y] = a_1\mu_1 + \cdots + a_n\mu_n$, $\sigma^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$.

- Linear combinations of iid normals is normal
  - Proof: Use moment generating functions

Gamma distribution

$$X \sim \text{Gamma}(\alpha, \beta) \text{ if PDF } f_X(x) \propto x^{\alpha-1}e^{-\beta x}, x > 0$$

- $\alpha > 0$: shape parameter, $\beta > 0$: rate parameter, $\theta = 1/\beta$: scale parameter
- Mean: α/β, Variance: α/β^2, MGF: $(1 - \lambda/\beta)^{-\alpha}$, $\lambda < \beta$
- Sum of n iid Exp(β) is Gamma(n, β)
  - Proof: Use MGF (mostly)
- Square of Normal($0, \sigma^2$) is Gamma($1/2, 1/2\sigma^2$)
  - Proof: Use CDF method

Cauchy distribution

$$X \sim \text{Cauchy}(\theta, \alpha^2) \text{ if PDF } f_X(x) = \frac{1}{\pi} \frac{\alpha}{\alpha^2 + (x-\theta)^2}$$

- θ: location parameter, $\alpha > 0$: scale parameter
- Mean: undefined, Variance: undefined, MGF: undefined
- Suppose $X, Y \sim$ iid Normal$(0, \sigma^2)$. Then,

$$\frac{X}{Y} \sim \text{Cauchy}(0, 1)$$

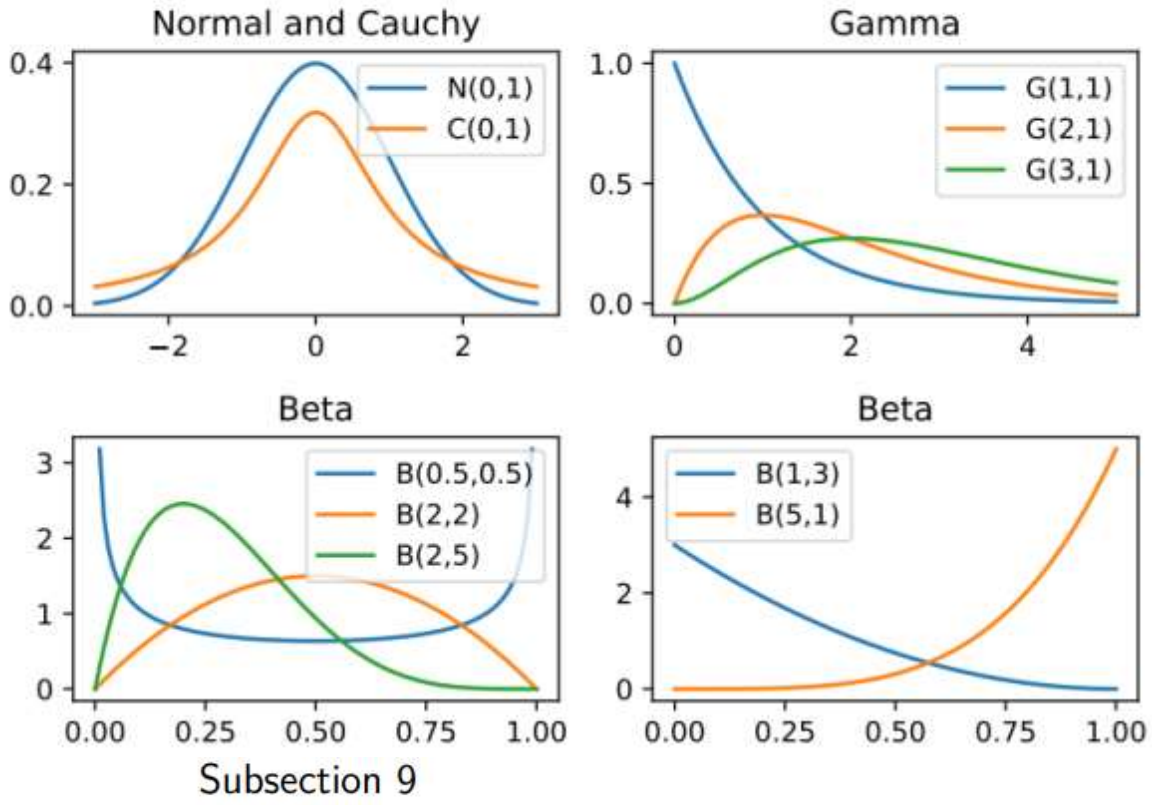
Beta distribution

$$X \sim \text{Beta}(\alpha, \beta) \text{ if PDF } f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1$$

- $\alpha > 0, \beta > 0$: shape parameters
- Mean: $\alpha/(\alpha+\beta)$, Variance: $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
- Beta$(\alpha, 1)$ has PDF $\propto x^{\alpha-1}$: power function distribution
- Suppose $X \sim \text{Gamma}(\alpha, 1/\theta)$, $Y \sim \text{Gamma}(\beta, 1/\theta)$, then

$$\frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)$$

Plots of PDFs



Descriptive statistics of normal samples

Normal samples

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- Very common assumption in many situations
  - CLT is used as justification, often
- Sample mean
$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$
- Sample variance
$$S^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$$
- Recall: Sample mean and sample variance are random variables
- For normal samples, the distribution of the sample mean and variance can be characterised in more detail

Distribution of Sample Mean

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- $\overline{X} = \frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n$
- Sample mean is a linear combination of iid normal random variables
  - So, Sample mean has a normal distribution

$$\overline{X} \sim \text{Normal}(\mu, \sigma^2/n)$$

- $E[\overline{X}] = \mu$
- $\text{Var}(\overline{X}) = \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 = \sigma^2/n$

Sum of squares of normal samples: Chi-square

$$X_1, \ldots, X_n \sim \text{iid Normal}(0, \sigma^2)$$

- X_i^2: Gamma$(1/2, 1/2\sigma^2)$, independent
- **Result**: Sum of n independent Gamma(α, β) is Gamma$(n\alpha, \beta)$

$$X_1^2 + \cdots + X_n^2 \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2\sigma^2})$$

- Gamma$(n/2, 1/2)$: called Chi-square distribution with n degrees of freedom, denoted χ^2_n

Sample mean and variance of normal samples

Theorem

Suppose $X_1, \ldots, X_n \overset{iid}{\sim}$ Normal(μ, σ^2). Then,

1. $\overline{X} \sim$ *Normal*$(\mu, \sigma^2/n)$
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, *Chi-square with $n-1$ degrees of freedom.*
3. $\overline{X}$ *and* S^2 *are independent.*

- For normal samples, the joint distribution of sample mean and variance is precisely known.

$$S^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$$

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{n}\right) = \overline{X} - \overline{X} = 0 \qquad \frac{(n-1)S^2}{\sigma^2} = \left(\frac{X_1 - \overline{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \overline{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

Parameter estimation

Section 1

Statistical problems in real life

Example 1: Who is the best captain in the IPL?

- Problem and planning
  - What are the qualities of a good captain? How to quantify it?
  - Typically, no unique or easy answer. How to approach?
- Data
  - What data is available? Scoresheets. How to collect it and consolidate?
  - Should some sampling of experts or fans be done?
- Analysis
  - Study data: descriptive stats, histograms, scatter plots
  - Find patterns and fit models or form hypotheses
    - Use statistical procedures to find unknown parameters or test hypothesis
  - Derive metrics that measure captaincy in the IPL
- Conclusion and communication
  - Develop visualizations for communicating results

Example 2: How many tigers are there in India?

- National Tiger Conservation Authority (NTCA)
  - Statutory body for strengthening tiger conservation [ntca.gov.in]
- Tiger census
  - Sampling over multiple phases/methods
    - ★ Survey by field forest staff
    - ★ Landscape characterization using satellite and other data
    - ★ Intensive camera traps
- Statistical methods
  - Find relationships between tiger population and various factors
  - Find a joint distribution likelihood model
  - Estimate number of tigers not camera-trapped

2018: 2461 tigers camera-trapped, 2967 total estimated tigers

Example 3: Was a remote-proctored exam successful?

- Problem and planning
  - How to assess success of exam?
  - Honor code, possible collaboration
- Data
  - Scores in online exam
  - Scores in previous in-person exams
- Analysis
  - Test the hypothesis that "honor code" was violated
  - Estimate number of violations
  - Detect violators or groups of collaborators
- Conclusion and communication
  - To university authorities
  - To students

The importance of communication

- Consider 1500 students in a course
  - 1 honor code violation in an in-person exam
  - 2 honor code violations in a remote-proctored exam
- Communication 1: **100% increase in honor code violations in remote-proctored exams**
- Communication 2: **Honor code violations within 0.15% under remote proctoring**
- You will see such communication in the press and in social media
- Truthful representation of what the data has conveyed is often difficult to find

Summary

- Statistical problems in real life
  - Usually very complex and involve several competing factors that are difficult to measure
  - Many such questions have very serious consequences
    - Medical problems, societal problems, policy and development issues
  - Majority of time is usually spent on planning, acquiring data and in formulating a clear communication plan
- In this course
  - We will focus on the "Analysis" part
  - Analysis: involves well-defined statistical procedures assuming an iid sample model for available data
  - Estimation of unknown parameters in
    - probabilistic models formulated for data
    - relationship models between factors
  - Testing of hypothesis using data

Books

- Our textbook for the course
- Mathematical Statistics and Data Analysis by John A. Rice, CENGAGE learning
  - Good reference with all theory and equations
  - Has lots of data and practical examples
- The Art of Statistics (Learning from Data) by David Spiegelhalter, Pelican
  - A popular book about the entire statistical approach to problem solving and understanding phenomenon
  - No equations
  - Emphasis on all aspects of the underlying problem

Section 2

Introduction to parameter estimation

Illustrative example 1: Bernoulli(p) trials

- Setting
  - n Bernoulli(p) trials with p unknown
  - One set of samples ($n = 10$): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - Can you guess p?
- Above is an example of a simple "parameter estimation" problem
  - Result of Bernoulli trial is a random variable: $X \sim \{0, 1\}$
    - Distribution of X: $\text{Prob}(X = 0) = 1 - p$, $\text{Prob}(X = 1) = p$
    - p: a parameter of the distribution
  - We observe a certain number of iid samples from the distribution
    - Using the observed samples, we are required to estimate a parameter

Illustrative example 2: Emissions of alpha particles

- Number of particles N emitted in a 10 sec interval
  - Modelled as Poisson: $P(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}$
  - λ is a parameter: average number of particles emitted
- Samples from one observation

n	Observed	Poisson fit	n	Observed	Poisson fit
0-2	18	12.2	10	123	130.6
3	28	27.0	11	101	99.7
4	56	56.5	12	74	69.7
5	105	94.9	13	53	45.0
6	126	132.7	14	23	27.0
7	146	132.7	15	15	15.1
8	164	166.9	16	9	7.9
9	161	155.6	17+	5	7.1

- Parameter esimation: What is λ?

Illustrative example 3: Noise in electronic circuits

- Voltage or current measured in circuits will show random fluctuations
  - The same voltage measured at different times will give slightly different values
- Popular model for such voltages or currents in circuits

$$X \sim \text{Normal}(\mu, \sigma^2)$$

- Two parameters
  - μ: a parameter (average voltage or current)
  - σ^2: another parameter (variance)
- 10 measurements: 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
- Parameter estimation: What is μ? What is σ?

Parameter estimation

- iid samples
$$X_1, \ldots, X_n \sim \text{iid } X$$
- X has a distribution described by some parameters θ_1, θ_2, ...
  - Parameters take real values, $\theta_i \in \mathbb{R}$
- Parameter estimation: What is θ_1? What is θ_2? $\cdots$
- Estimator for a parameter θ
  - Function of the samples: $\hat{\theta}(X_1, \ldots, X_n)$
  - Notation: $\hat{\theta}$ is an estimator for parameter θ
- Parameter vs estimator
  - θ: constant parameter, not a random variable
  - $\hat{\theta}$: function of n random variables; therefore, it is a random variable
    - $\hat{\theta}$ will have a distribution, PMF or PDF

Example: Bernoulli(p) trials

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Parameter: p
  - We had denoted this as θ earlier
- Estimator 1: $\hat{p}_1 = 1/2$
- Estimator 2: $\hat{p}_2 = (X_1 + X_2)/2$
- Estimator 3: $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$
- An infinite number of estimators are possible
  - How to characterize *good* estimators?
  - How to design estimators?

Section 3

Error in estimation

Estimation error

- θ: parameter, $\hat{\theta}(X_1, \ldots, X_n)$: estimator
  - Error: $\hat{\theta}(X_1, \ldots, X_n) - \theta$ is a random variable
- We expect the estimator random variable $\hat{\theta}(X_1, \ldots, X_n)$ to take values around the actual value of the parameter θ. So, the random variable 'Error' should take values close to 0.
  - How to express this mathematically? $P(|\text{Error}| > \delta)$ should be small
- Parameter will be in a certain range, and estimator error should be low over the entire range
  - How to quantify 'low'?
  - Example: In Bernoulli(p) trials, $p \in [0, 1]$, and the same estimator has to give low error for all values of p
    - What is 'low'? |Error| should be small compared to p.
    - 10% or lower error: $|\text{Error}| \leq p/10$

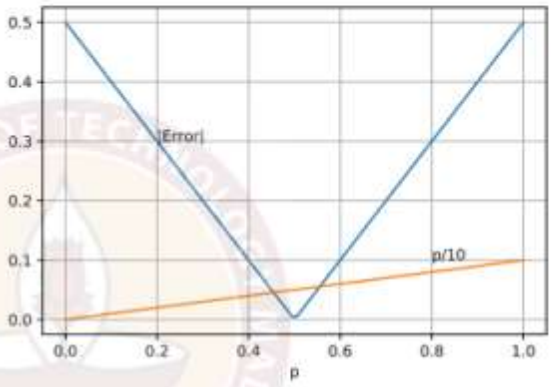
Example: Bernoulli(p) trials

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$, $\hat{p}_2 = (X_1 + X_2)/2$, $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$
- Variation in samples => variation in estimation
  - 10 samples of Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
    - $\hat{p}_1 = 0.5$, $\hat{p}_2 = 0.5$, $\hat{p}_3 = 0.5$
  - 10 samples in another round: 1, 0, 0, 1, 0, 1, 0, 1, 0, 0
    - $\hat{p}_1 = 0.5$, $\hat{p}_2 = 0.5$, $\hat{p}_3 = 0.4$
  - 10 samples in another round: 1, 1, 0, 0, 0, 1, 0, 1, 0, 1
    - $\hat{p}_1 = 0.5$, $\hat{p}_2 = 1$, $\hat{p}_3 = 0.5$
- $\hat{p}_1$: does not work for all values of p
- $\hat{p}_2$: varies a lot with variation in samples
- $\hat{p}_3$: seems to be promising

Example: Bernoulli(p) trials (contd)

- $\hat{p}_1 = 1/2$
  - Error: $1/2 - p$
- $P(|\text{Error}| > p/10) = 1$ if $p < 5/11$ or $p > 5/9$



- $\hat{p}_2 = (X_1 + X_2)/2$
  - Error: $\frac{X_1+X_2}{2} - p$
- $P(|\text{Error}| > p/10) = 1$ if $p < 5/11$ or $5/9 < p < 10/11$

x_1	x_2	$e = \frac{x_1+x_2}{2} - p$	$\Pr(\text{Error} = e)$
0	0	$-p$	$(1-p)^2$
0	1	$1/2 - p$	$p(1-p)$
1	0	$1/2 - p$	$p(1-p)$
1	1	$1 - p$	p^2

Example: Bernoulli(p) trials (contd)

- $\hat{p}_3 = (X_1 + \cdots + X_n)/n$
  - Error: $(X_1 + \cdots + X_n)/n - p$
- Recall: Chebyshev bound

$$P(|\text{Error} - E[\text{Error}]| > \delta) \leq \frac{\text{Var}(\text{Error})}{\delta^2}$$

- Using above bound,

$$P(|\text{Error}| > p/10) \leq \frac{p(1-p)/n}{p^2/100} \leq \frac{100(1-p)}{np}$$

- For any fixed p, the above probability tends to 0 as $n \to \infty$
  - Chebyshev bound results in fall of $1/n$
  - Use Chernoff bound or concentration to get exponential fall with n

Observations

- Various estimators are usually possible
  - Every estimator will have an error and the error will have a distribution
- Bounds on $P(|\text{Error}| > \delta)$ are interesting and capture useful properties of the estimator
  - Good design: $P(|\text{Error}| > \delta)$ will fall with n
- Chebyshev bound is a useful tool

$$P(|\text{Error} - E[\text{Error}]| > \delta) \leq \frac{\text{Var(Error)}}{\delta^2}$$

- Good design principles
  - $E[\text{Error}]$ should be close to or equal to 0
  - Var(Error)$\rightarrow$ 0 with n

Section 4

Bias, Variance and Risk of an Estimator

Recap: Mean and variance

- Random variable X taking values in a set $\mathcal{X}$
  - Assume discrete with a PMF f_X
- Mean or expected value of X
  - $E[X] = \sum_{x \in \mathcal{X}} x f_X(x)$
  - Average value
  - Denoted μ

Continuous with PDF $f_X(x)$

$E[X] = \int x f_X(x)\, dx$

$E[X^2] = \int x^2 f_X(x)\, dx$

- Second moment: $E[X^2] = \sum_{x \in \mathcal{X}} x^2 f_X(x)$
- Variance: $\text{Var}(X) = E[(X - \mu)^2]$
  - $\text{Var}(X) = E[X^2] - \mu^2$ or $E[X^2] = \text{Var}(X) + \mu^2$
  - Spread of the distribution
- Low variance: random variable takes values around μ

Bias

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Definition (Bias)

The bias of the estimator $\hat{\theta}$ for a parameter θ, denoted $\text{Bias}(\hat{\theta}, \theta)$ is defined as

$$\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta.$$

- Since Error $= \hat{\theta} - \theta$, bias is the expected value of Error
- An estimator with bias equal to 0 is said to be an *unbiased* estimator

Risk (squared error)

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Definition (Risk)

The (squared-error) risk of the estimator $\hat{\theta}$ for a parameter θ, denoted $\text{Risk}(\hat{\theta}, \theta)$, is defined as

$$\text{Risk}(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2].$$

- Since Error $= \hat{\theta} - \theta$, risk is the expected value of "squared error" and is also called mean squared error (MSE) often
- Squared-error risk is the second moment of Error

Variance

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short
- Variance of estimator

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

- Variance of error: $\text{Error} = \hat{\theta} - \theta$
  - Error is a "translated" version of the estimator $\hat{\theta}$
  - Remember: θ is a constant

$$\text{Var}(\text{Error}) = \text{Var}(\hat{\theta})$$

Bias-variance tradeoff

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Theorem (Bias-variance tradeoff)

The risk of the estimator satisfies the following relationship:

$$Risk(\hat{\theta}, \theta) = Bias(\hat{\theta}, \theta)^2 + Var(\hat{\theta})$$

- Expanded form

$$E[(\hat{\theta} - \theta)^2] = E[\hat{\theta} - \theta]^2 + E[(\hat{\theta} - E[\hat{\theta}])^2]$$

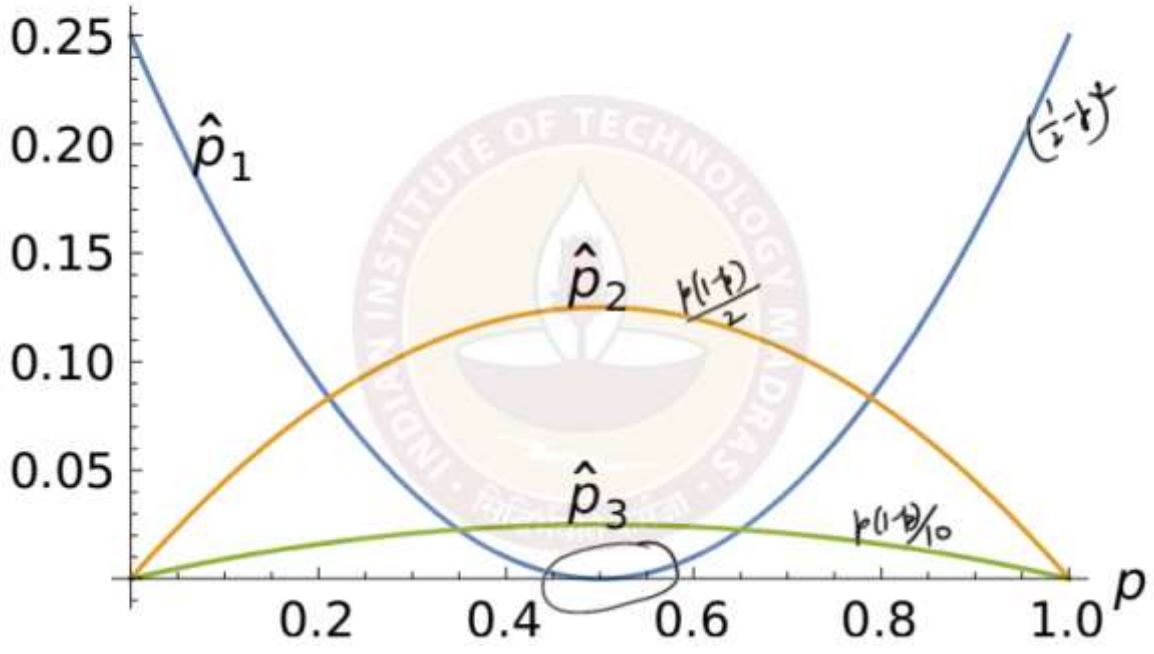
- Proof
  - $\text{Risk} = \text{E}[\text{Error}^2] = \text{Mean}[\text{Error}]^2 + \text{Var}[\text{Error}]$

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$
  - Bias $= 1/2 - p$, Variance $= 0$, Risk $= (1/2 - p)^2$
- $\hat{p}_2 = (X_1 + X_2)/2$
  - Bias $= 0$
  - Variance $= \frac{1}{4}(\text{Var}(X_1) + \text{Var}(X_2)) = p(1-p)/2$
  - Risk $= p(1-p)/2$
- $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$
  - Bias $= 0$
  - Variance $= \frac{1}{n^2}(\text{Var}(X_1) + \cdots + \text{Var}(X_n)) = p(1-p)/n$
  - Risk $= p(1-p)/n$ → falling with n

Plot of Risk versus p, $n = 10$



Problem: Computing bias, variance, risk

Let $X_1, \ldots, X_n \sim$ iid Bernoulli(p). Consider the estimator

$$\hat{p} = \frac{X_1 + \cdots + X_n + \sqrt{n}/2}{n + \sqrt{n}}.$$

Find the bias, variance and risk of $\hat{p}$.

Section 5

Estimator design approach: Method of moments

Moments and parameters

$$X \sim f_X(x), \text{ parameters } \theta_1, \theta_2, \ldots$$

- Moments $E[X]$, $E[X^2]$, etc can be expressed as functions of the parameters
- Bernoulli(p)
  - $E[X] = p$
- Poisson(λ)
  - $E[X] = \lambda$
- Exponential(λ)
  - $E[X] = 1/\lambda$
- Normal(μ, σ^2)
  - $E[X] = \mu$, $E[X^2] = \mu^2 + \sigma^2$
- Gamma(α, β)
  - $E[X] = \alpha/\beta$,
    $E[X^2] = \alpha^2/\beta^2 + \alpha/\beta^2$

Moments of samples

$$X_1, \ldots, X_n \sim \text{iid } X$$

- Sample moments

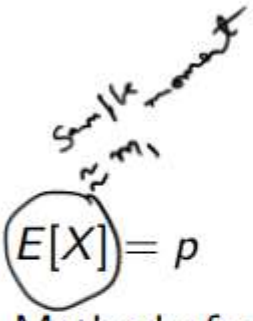
$$M_k(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

- One sampling instance: $x_1, \ldots, x_n$
  - 1st sample moment: $m_1 = \frac{1}{n}(x_1 + \cdots + x_n)$
  - k-th sample moment: $m_k = \frac{1}{n}(x_1^k + \cdots + x_n^k)$
- M_k is a random variable, and m_k is the value taken by it in one sampling instance
  - If sampling is repeated, the random variable M_k will take different values
  - We expect that M_k will take values around $E[X^k]$
    - ⋆ Justified by WLLN, CLT, concentration

Method of moments

- Procedure
  - Equate sample moments to expression for moments in terms of unknown parameters
  - Solve for the unknown parameters
- One parameter θ usually needs one moment
  - Sample moment: m_1
  - Distribution moment: $E[X] = f(\theta)$
  - Solve for θ from $f(\theta) = m_1$ in terms of m_1
  - $\hat{\theta}$: replace m_1 by M_1 in above solution
- Two parameters θ_1, θ_2 usually needs two moments
  - Sample moments: m_1, m_2
  - Distribution moments: $E[X] = f(\theta_1, \theta_2)$, $E[X^2] = g(\theta_1, \theta_2)$
  - Solve for θ_1, θ_2 from $f(\theta_1, \theta_2) = m_1$, $g(\theta_1, \theta_2) = m_2$ in terms of m_1, m_2
  - $\hat{\theta}_1, \hat{\theta}_2$: replace m_1 by M_1 and m_2 by M_2 in above solution

Example: Bernoulli(p)



$X_1, \ldots, X_n \sim$ iid Bernoulli(p)

- $E[X] = p$
- Method of moments equation

$$p = m_1$$

- Estimator

$$\hat{p} = M_1 = \frac{1}{n}(X_1 + \cdots + X_n)$$

Example: Poisson

$X_1, \ldots, X_n \sim$ iid Poisson(λ)

- $E[X] = \lambda$
- Method of moments equation

$$\lambda = m_1$$

- Estimator

$$\hat{\lambda} = M_1 = \frac{1}{n}(X_1 + \cdots + X_n)$$

Example: Exponential

$$X_1, \ldots, X_n \sim \text{iid Exp}(\lambda)$$

- $E[X] = 1/\lambda$
- Method of moments equation: $1/\lambda = m_1$
- Solution: $\lambda = 1/m_1$
- Estimator

$$\hat{\lambda} = 1/M_1 = \frac{n}{X_1 + \cdots + X_n}$$

Example: Normal

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- $E[X] = \mu$, $E[X^2] = \mu^2 + \sigma^2$
- Method of moments equation: $\mu = m_1$, $\mu^2 + \sigma^2 = m_2$
- Solution: $\mu = m_1$, $\sigma = \sqrt{m_2 - m_1^2}$
- Estimator for μ

$$\hat{\mu} = M_1 = \frac{X_1 + \cdots + X_n}{n}$$

- Estimator for σ

$$\hat{\sigma} = \sqrt{\frac{X_1^2 + \cdots + X_n^2}{n} - \frac{(X_1 + \cdots + X_n)^2}{n^2}}$$

Problem: Gamma

$$X_1, \ldots, X_n \sim \text{iid Gamma}(\alpha, \beta)$$

- $E[X] = \alpha/\beta$, $E[X^2] = \alpha^2/\beta^2 + \alpha/\beta^2$

.....

Problem: Binomial(N, p)

$$X_1, \ldots, X_n \sim \text{iid Binomial}(N, p)$$

- $E[X] = Np$, $E[X^2] = N^2p^2 + Np(1-p)$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$
- Alpha particles emission in 10sec: Poisson(λ)
  - No of particles emitted per second $= 0.8392$
  - $\hat{\lambda}$ = Average number of particles emitted in 10 seconds $= 8.392$
- Normal(μ, σ^2): 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
  - $\hat{\mu} = m_1 = (1.07 + 0.91 + \cdots + 1.08)/10 = 1.024$
  - $\hat{\sigma} = \sqrt{m_2 - m_1^2} = \sqrt{1.05482 - 1.024^2} = 0.079$
- Binomial(N, p): 8, 7, 6, 11, 8, 5, 3, 7, 6, 9
  - $\hat{N} = 19$
  - $\hat{p} = 0.371$

Section 6

Estimator design approach: Maximum likelihood

Likelihood of *iid* samples

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- $f_X(x)$: depends on $\theta_1, \theta_2, \ldots$
  - To bring this out, we will write $f_X(x)$ as $f_X(x; \theta_1, \theta_2, \ldots)$
  - Example: Normal(μ, σ^2)
    - ⋆ $f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$
- Likelihood of a sampling $x_1, x_2, \ldots, x_n$, denoted $L(x_1, \ldots, x_n)$

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $L = p(1-p)(1-p)p(1-p)ppp(1-p)(1-p) = p^5(1-p)^5$
- Normal(μ, σ^2): 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
  - $L = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1.07-\mu)^2/2\sigma^2} \cdots \frac{1}{\sqrt{2\pi}\sigma} e^{-(1.08-\mu)^2/2\sigma^2}$
  - Simplified: $\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{10} e^{-((1.07-\mu)^2+\cdots+(1.08-\mu)^2)/2\sigma^2}$

Maximum likelihood (ML) estimator

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- Likelihood $L(x_1, \ldots, x_n)$ is a function of parameters
- Maximum likelihood (ML) estimation
  - Sampling: $x_1, \ldots, x_n$

$$\theta_1^*, \theta_2^*, \ldots = \arg\max_{\theta_1, \theta_2} \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

- Find parameters that maximize likelihood for a given set of samples
- When the maximization problem has a closed-form solution, the estimator can be expressed in terms of the samples.
  - This will need a lot of algebraic manipulation.
- In many cases, the maximization problem will need a numerical routine.
  - Several standard modules are available for optimization.

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Samples: $x_1, x_2, \ldots, x_n$
  - $x_i = 0$ or 1
- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$
  - $f_X(x_i) = p$ if $x_i = 1$, or $f_X(x_i) = 1 - p$ if $x_i = 0$
  - Let w denote the number of 1s in the sample

$$L(x_1, \ldots, x_n) = p^w (1-p)^{n-w}$$

- ML estimation: $p^* = \arg\max_p p^w (1-p)^{n-w}$
  - How to find the p that maximizes the above expression?
  - Differentiate w.r.t. p and equate to 0 and solve for p
  - $p^* = w/n = \frac{x_1 + \cdots + x_n}{n}$

replace x_i by X_i

$$\hat{p}_{ML} = \frac{X_1 + \cdots + X_n}{n}$$

Same as MME (method of moments estimator)

Example: Poisson(λ)

$$X_1, \ldots, X_n \sim \text{iid Poisson}(\lambda)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i-1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$

$$L(x_1, \ldots, x_n) = \frac{1}{x_1! \cdots x_n!} e^{-n\lambda} \lambda^{x_1 + \cdots + x_n}$$

- ML estimation: $\lambda^* = \arg\max_{\lambda} [(x_1 + \cdots + x_n) \log \lambda - n\lambda]$
  - $\lambda^* = \frac{x_1 + \cdots + x_n}{n}$

$$\hat{\lambda}_{ML} = \frac{X_1 + \cdots + X_n}{n}$$ ← sample mean ← same as MME

Example: Normal(μ, σ^2)

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

$$L(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2}$$

$x_i^2 - 2\mu x_i + \mu^2$

- ML estimation: $\mu^*, \sigma^* = \arg\min_{\mu, \sigma} \left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + n \log \sigma\right]$

Sample mean same as MME ← $\hat{\mu}_{ML} = \frac{X_1 + \cdots + X_n}{n}$ ← diff. w.r.t. μ (treat σ as constant)

$$\hat{\sigma}^2_{ML} = \frac{(X_1 - \hat{\mu}_{ML})^2 + \cdots + (X_n - \hat{\mu}_{ML})^2}{n}$$ ← diff. w.r.t. σ (treat μ as constant)

Sample variance

Observations

- Maximum likelihood is a very popular method for deriving estimators
- Theoretically and intuitively appealing: maximize the probability or likelihood of the observed samples
- Deriving the actual estimator needs some careful calculus
- Numerous questions
  - How do ML estimators look? They seem similar to MME, so far.
  - How does MME compare with ML? How to compare estimators?

Section 7

Finding MME and ML estimators
Problem: Exp(λ)

$$X_1, \ldots, X_n \sim \text{iid Exp}(\lambda)$$

Problem: $\{1, 2, 3\}$ w.p. p_1, p_2, p_3

parameters

$$X_1, \ldots, X_n \sim \text{iid } \{\overset{p_1}{1}, \overset{p_2}{2}, \overset{p_3}{3}\}$$

Problem: Uniform$[0, \theta]$

$$X_1, \ldots, X_n \sim \text{iid Uniform}[0, \theta]$$

Problem: Uniform$\{1, 2, \ldots, N\}$

$$X_1, \ldots, X_n \sim \text{iid Uniform}\{1, 2, \ldots, N\}$$

Problem: Gamma(α, β)

$$X_1, \ldots, X_n \sim \text{iid Gamma}(\alpha, \beta), f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Problem: Binomial(N, p)

$$X_1, \ldots, X_n \sim \text{iid Binomial}(N, p)$$

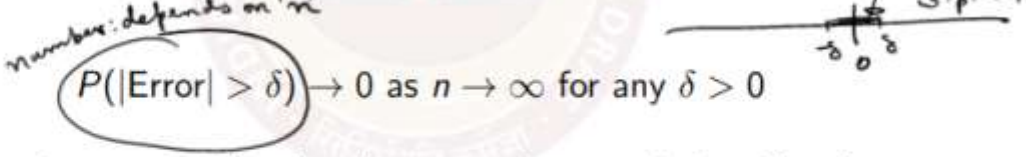
Section 8

Properties of Estimators

Consistency of estimators

$$X_1, \ldots, X_n \sim \text{iid } f_X(x; \theta)$$

- Estimator: $\hat{\theta}$, Error $= \hat{\theta} - \theta$
- 'Error' is a random variable
- As n increases, we expect 'Error' to take values that are close to zero. How to capture this requirement mathematically?



- If an estimator satisfies the above requirement, it is said to be *consistent*
- Technically, the above requirement is called *convergence in probability*

Examples: bias and consistency

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter: } \mu = E[X]$$

- Estimator 1: $\hat{\mu}_1 = M_1 = (X_1 + \cdots + X_n)/n$
  - This estimator is unbiased, $E[\hat{\mu}] = \mu$
  - This estimator is consistent (Proof: WLLN)
- Estimator 2: $\hat{\mu}_2 = M_1 = (X_1 + \cdots + X_n)/(n-1)$
  - This estimator is biased, $E[\hat{\mu}] = n\mu/(n-1) \neq \mu$
  - This estimator is consistent (Proof: $\hat{\mu}_2 = \hat{\mu}_1(1 - 1/n)$
- Estimator 3: $\hat{\mu}_3 = X_1$
  - This estimator is unbiased, $E[\hat{\mu}] = \mu$
  - This estimator is inconsistent (for most non-trivial distributions)

Estimator designs, properties and comparisons

- Method of moments estimators
  - If parameter is mean or variance, they will be unbiased
  - For most other cases, they may be biased
  - Usually, consistent
    - Sample moments converge to distribution moments
    - If MME estimate is a continuous function of moments, then the estimate converges too.
- Maximum likelihood estimators
  - Consistent
  - Bias vanishes in a limiting sense with growing n
  - Several interesting properties
    - Functional invariance: ML estimate of $g(\theta)$ is $g(\hat{\theta}_{ML})$ for smooth functions g
- How to compare estimators? Squared-error risk or Mean Squared Error (MSE) is one option

MSE or Risk of estimators

- Finding the risk of an estimator theoretically usually involves some calculations of expectations
- Example: $X_1, \ldots, X_n \sim$ iid Uniform$[0, \theta]$
  - $\hat{\theta}_{MME} = 2M_1$
    - ⋆ Bias $= 0$, Risk $=$ Variance
    - ⋆ Risk$(\hat{\theta}_{MME}) = \dfrac{\theta^2}{3n}$
  - $\hat{\theta}_{ML} = \max(X_1, \ldots, X_n)$
    - ⋆ $f_{\hat{\theta}}(t) = \frac{nt^{n-1}}{\theta^n}$, $E[\hat{\theta}] = \frac{n\theta}{n+1}$, $E[\hat{\theta}^2] = \frac{n\theta^2}{n+2}$, $\text{Var}(\hat{\theta}) = \frac{n\theta^2}{(n+2)(n+1)^2}$
    - ⋆ Bias $= -\theta/(n+1)$, Risk $= \dfrac{2\theta^2}{(n+1)(n+2)} \leq \dfrac{2\theta^2}{n^2}$
  - ML is a factor of $1/n$ better than MME!
- A good alternative to theoretical computations is Monte Carlo simulations, which will work in most cases
  - Colab exercise: build a simulation for above and show ML estimator's risk is $1/n$ better than MME

Section 9

Confidence intervals
Example: Surveys and results

- "82% Indians willing to take Covid-19 vaccine," from a *Gallup Survey*
  - **3045** people were called and asked if they were willing to take a vaccine between **Nov 24, 2020 and Jan 8, 2021**
  - Languages spoken: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Odia, Punjabi, Assamese
  - 95% confidence level with a margin of error of 3%
- "40% Indians willing to take vaccine," from a *LocalCircles survey*
  - **9628** votes on **Jan 25, 2021** through the LocalCircles app
  - Demographics: people from 299 districts, 48% tier 1, 27% tier 2, 25% rural

This lecture: What are confidence level and margin of error?

Estimation of sample mean and confidence interval

$$X_1, \ldots, X_n \sim \text{iid } X, \mu = E[X]$$

- Estimator: $\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$
- Suppose $\Pr(|\hat{\mu} - \mu| < 0.03) = 0.95$.
  - Probability that μ lies in the interval $[\hat{\mu} - 0.03, \hat{\mu} + 0.03]$ is 0.95
  - $[\hat{\mu} - 0.03, \hat{\mu} + 0.03]$: called *95%-confidence interval*
  - $\hat{\mu}$ in one sampling instance: estimate with margin of error 3% at confidence level 95%
- **Confidence interval** (in general)
  - Suppose $\Pr(|\hat{\mu} - \mu| < \alpha) = \beta$, where α is a small fraction and β is a large fraction
  - $\hat{\mu}$ in one sampling instance: estimate with **margin of error** $(100\alpha)\%$ at confidence level $(100\beta)\%$

How to find α, β for which $\Pr(|X - \mu| < \alpha) = \beta$?

- Suppose X is continuous and has CDF F_X
  - $P(X \le x) = F_X(x)$

$$P(|X - \mu| < \alpha) = P(\mu - \alpha < X < \mu + \alpha) = F_X(\mu + \alpha) - F_X(\mu - \alpha)$$

- Given β, find α such that $F_X(\mu + \alpha) - F_X(\mu - \alpha) = \beta$
- Suppse X is symmetric about the mean,
  i.e. $P(X < \mu - \alpha) = P(X > \mu + \alpha)$
  - $F_X(\mu + \alpha) = 1 - P(X > \mu + \alpha) = 1 - P(X < \mu - \alpha) = 1 - F_X(\mu - \alpha)$
  - $F_X(\mu + \alpha) - F_X(\mu - \alpha) = 1 - 2F_X(\mu - \alpha)$
  - Given β, find α s.t. $1 - 2F_X(\mu - \alpha) = \beta$ or $F_X(\mu - \alpha) = (1 - \beta)/2$

Normal samples with known variance

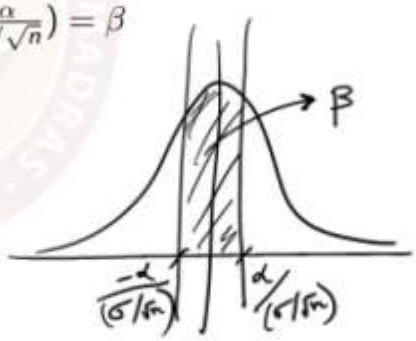
$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2), \sigma^2 \text{ known}$$

- Estimator: $\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$
- $\hat{\mu} \sim \text{Normal}(\mu, \sigma^2/n)$, $Z = \dfrac{\hat{\mu} - \mu}{(\sigma/\sqrt{n})} \sim \text{Normal}(0,1)$
  - $P(|\hat{\mu} - \mu| < \alpha) = \beta \leftrightarrow P(\left|\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}\right| < \frac{\alpha}{\sigma/\sqrt{n}}) = \beta$
  - $\leftrightarrow P(|\text{Normal}(0,1)| < \frac{\alpha}{\sigma/\sqrt{n}}) = \beta.$

Standard: its CDF is well-known

β	$\frac{\alpha}{\sigma/\sqrt{n}}$
0.68	0.99
0.90	1.64
0.95	1.96

Using CDF of Standard normal



Problem: Normal samples

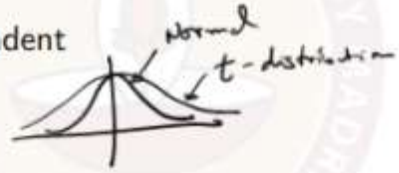
Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and $\sigma = 3$. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\sigma = 3$
- $\beta = 0.95$, and, using the CDF of Normal(0,1), we get $\frac{\alpha}{\sigma/\sqrt{n}} = 1.96$
- $\alpha = 1.96 \times 3/\sqrt{16} = 1.47$
  - $P(|\hat{\mu} - \mu| < 1.47) = 0.95$
- 95% confidence interval: $[10.06 - 1.47, 10.06 + 1.47] = [8.59, 11.53]$

Normal samples and t-distribution

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
  - $\overline{X} \sim \text{Normal}(0, \sigma^2/n)$, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
  - $\overline{X}$ and S are independent
- $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$



- $\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim$ t-distribution with $n-1$ degrees of freedom, denoted t_{n-1}
  - PDF of t_n: proportional to $(1 + x^2/n)^{-(n+1)/2}$
  - Assume that CDF of t_n is known in calculations. Computer packages can provide CDF.

Normal samples with unknown variance

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2), \sigma^2 \text{ unknown}$$

- Sample instance: $x_1, \ldots, x_n$
- Estimated mean and variance: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

$\hat{\mu} = (x_1 + \cdots + x_n)/n$

- $\hat{\mu} \sim \text{Normal}(\mu, \sigma^2/n)$, $Z = \frac{\hat{\mu} - \mu}{(S/\sqrt{n})} \sim t_{n-1}$
  - $P(|\hat{\mu} - \mu| < \alpha) = \beta$ approx. $P(\left|\frac{\hat{\mu} - \mu}{S/\sqrt{n}}\right| < \frac{\alpha}{\hat{\sigma}/\sqrt{n}}) = \beta$
  - $\leftrightarrow P(|t_{n-1}| < \frac{\alpha}{\hat{\sigma}/\sqrt{n}}) = \beta$.

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and σ unknown. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\hat{\sigma} = 3.297$
- $\beta = 0.95$ and, using CDF of t_{15}, we get $\frac{\alpha}{\hat{\sigma}/\sqrt{n}} = 2.13$
- $\alpha = 2.13 \times 3.297/\sqrt{16} = 1.76$
  - $P(|\hat{\mu} - \mu| < 1.76) \approx 0.95$
- 95% confidence interval: $[10.06 - 1.76, 10.06 + 1.76] = [8.30, 11.82]$

What if samples are not normal?

- Use CLT to argue that sample mean will have a normal distribution
  - Use the same procedure as for normal, if reasonable
- If the specific distribution is known, an expression or bound for the sample variance may be possible
- Bernoulli(p) samples
  - This is common in most sampling surveys
    - ★ Response is either yes or no
  - Sample variance: $\frac{p(1-p)}{n} \leq \frac{0.25}{n}$
    - ★ $\frac{\hat{\sigma}}{\sqrt{n}} \approx \sqrt{\frac{0.25}{n}}$ is commonly used
    - ★ 95% confidence interval: $[\hat{\mu} - 1.96\sqrt{\frac{0.25}{n}}, \hat{\mu} + 1.96\sqrt{\frac{0.25}{n}}]$

Revisit survey example

- "82% Indians willing to take Covid-19 vaccine," from a *Gallup Survey*
  - **3045** people were called and asked if they were willing to take a vaccine between **Nov 24, 2020 and Jan 8, 2021**
  - Languages spoken: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Odia, Punjabi, Assamese
  - 95% confidence level with a margin of error of 3%
- "40% Indians willing to take vaccine," from a *LocalCircles survey*
  - **9628** votes on **Jan 25, 2021** through the LocalCircles app
  - Demographics: people from 299 districts, 48% tier 1, 27% tier 2, 25% rural
  - 95% confidence interval will have a lower margin of error

Most probably, the two surveys are sampling different types of populations!

Bayesian estimation and hypothesis testing

Section 1

Bayesian estimation
Parameter estimation

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Two schools of thought for design of estimators
- Frequentist: treat θ as an unknown constant
  - Method of moments
  - Maximum likelihood
- Bayesian: treat θ as a random variable with a known distribution
  - Bayesian estimation

Example 1: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Suppose that $p \sim \text{Uniform}\{0.25, 0.75\}$
  - Assume p is chosen first at random according to the above distribution
  - Once p is chosen, the samples are drawn according to Bernoulli(p)
- Samples: $1, 0, 1, 1, 0$
  - Notation: $S = (X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0)$
  - Estimate using Bayes' rule
    - $P(p = 0.25|S) = P(S|p = 0.25)P(p = 0.25)/P(S) = 0.25^3 \times 0.75^2 \times 0.5/P(S) = 0.25$
    - $P(p = 0.75|S) = 0.75^3 \times 0.25^2 \times 0.5/P(S) = 0.75$
    - $P(S) = 0.25^3 \times 0.75^2 \times 0.5 + 0.75^3 \times 0.25^2 \times 0.5 = 0.25^2 \times 0.75^2 \times 0.5$
  - Estimator 1: Since $P(p = 0.75|S) > P(p = 0.25|S)$, we could estimate $\hat{p} = 0.75$
  - Estimator 2: Posterior mean, $\hat{p} = 0.25\, P(p = 0.25|S) + 0.75\, P(p = 0.75|S) = 0.625$

Example 2: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Suppose that $p \sim \{\overset{0.9}{0.25}, \overset{0.1}{0.75}\}$
- Samples: $1, 0, 1, 1, 0$
  - Notation: $S = (X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0)$
  - Estimate using Bayes' rule
    - $P(p = 0.25|S) = P(S|p = 0.25)P(p = 0.25)/P(S) = 0.25^3 \times 0.75^2 \times 0.9/P(S) = 0.75$
    - $P(p = 0.75|S) = 0.75^3 \times 0.25^2 \times 0.1/P(S) = 0.25$
    - $P(S) = 0.25^3 \times 0.75^2 \times 0.9 + 0.75^3 \times 0.25^2 \times 0.1 = 0.25^2 \times 0.75^2 \times 0.3$
  - Estimator 1: Since $P(p = 0.25|S) > P(p = 0.75|S)$, we estimate $\hat{p} = 0.25$
  - Estimator 2: Posterior mean, $\hat{p} = 0.25\, P(p = 0.25|S) + 0.75\, P(p = 0.75|S) = 0.375$

Bayesian estimation

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \Theta$$

- Prior distribution of Θ: $\Theta \sim f_\Theta(\theta)$
- Samples: $x_1, \ldots, x_n$, Notation: $S = (X_1 = x_1, \ldots, X_n = x_n)$
- Bayes' rule: posterior $\propto$ likelihood $\times$ prior

$$P(\Theta = \theta | S) = P(S|\Theta = \theta) f_\Theta(\theta) / P(S)$$

- Estimation using "posterior" probability
  - Posterior mode: $\hat{\theta} = \arg\max_\theta P(S|\Theta = \theta) f_\Theta(\theta)$
  - Posterior mean: $\hat{\theta} = E[\Theta|S]$, mean of posterior distribution
    - $(\Theta|S)$ may be a known distribution, and its mean might become a simple formula in some cases

Meaning of prior distribution

- Prior distribution
  - Captures what we might know about the parameter
  - This could be using some scientific model or expert opinion
- Posterior $\propto$ Likelihood of samples $\times$ Prior
  - Intuitively understood as incorporating "data" into prior
  - Useful in modeling
- What if we do not know anything?
  - You can choose a flat prior, uniform over the entire range
- Lots of debates between frequentists vs Bayesians
  - Search "frequestist vs Bayesian"

Section 2

Choice of prior and examples

How to pick prior?

- Flat, uninformative
  - Nearly flat over the interval in which the parameter takes value
  - This usually reduces to something close to maximum likelihood
- Conjugate priors
  - Pick a prior so that the posterior is in the same class as prior
  - Examples
    - ⋆ Prior: Normal and Posterior: Normal
    - ⋆ Prior: Beta and Posterior: Beta
- Informative priors
  - This needs some justification from the domain of the problem
  - Parameterize the prior so that its flatness can be controlled

Bernoulli(p) samples with uniform prior

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(\mathbf{p})$$

- Prior $\mathbf{p} \sim \text{Uniform}[0, 1]$, continuous distribution
- Samples: $x_1, \ldots, x_n$
- Posterior: $\mathbf{p}|(X_1 = x_1, \ldots, X_n = x_n)$ is continuous
  - Posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n|\mathbf{p} = p) f_{\mathbf{p}}(p)$
  - Posterior density $\propto p^w (1-p)^{n-w}$, $0 \leq p \leq 1$
    - ⋆ $w = x_1 + \cdots + x_n$: number of 1s in samples
- Posterior density: $\text{Beta}(w+1, n-w+1)$
  - Posterior mean $= \frac{w+1}{w+1+n-w+1} = \frac{w+1}{n+2} = \frac{x_1+\cdots+x_n+1}{n+2}$

$$\hat{p} = \frac{X_1 + \cdots + X_n + 1}{n+2}$$

$\hat{p}_{ML} = \frac{X_1 + \cdots + X_n}{n}$

Bernoulli(p) samples with beta prior

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(\mathbf{p})$$

- Prior $\mathbf{p} \sim \text{Beta}(\alpha, \beta)$, continuous distribution
  - $f_{\mathbf{p}}(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$, $0 \le p \le 1$
- Samples: $x_1, \ldots, x_n$
- Posterior: $\mathbf{p} | (X_1 = x_1, \ldots, X_n = x_n)$ is continuous
  - Posterior density $\propto P(X_1 = x_1, \ldots, X_n = x_n | \mathbf{p} = p) f_{\mathbf{p}}(p)$
  - Posterior density $\propto p^{w+\alpha-1}(1-p)^{n-w+\beta-1}$, $0 \le p \le 1$
    - $w = x_1 + \cdots + x_n$: number of 1s in samples
- Posterior density: $\text{Beta}(w+\alpha, n-w+\beta)$
  - Posterior mean $= \frac{w+\alpha}{w+\alpha+n-w+\beta} = \frac{w+\alpha}{n+\alpha+\beta} = \frac{x_1+\cdots+x_n+\alpha}{n+\alpha+\beta}$

$$\hat{p} = \frac{X_1 + \cdots + X_n + \alpha}{n + \beta}$$

Observations for Beta prior

- Prior: $\text{Beta}(\alpha, \beta)$
  - $\alpha, \beta \ge 0$
  - PDF $\propto p^{\alpha-1}(1-p)^{\beta-1}$, $0 < p < 1$
  - How to pick α, ~~βeta~~?
- $\alpha = \beta = 1$: Uniform$[0, 1]$
  - Flat prior
  - Estimate close to, but not equal to, Maximum-Likelihood
- $\alpha = \beta = 0$
  - Estimate coincides with Maximum-Likelihood
- $\alpha = \beta$
  - Symmetric prior
- α, β may depend on n the number of samples
  - $\alpha = \beta = \sqrt{n}/2$ is an *interesting* choice

Normal samples with unknown mean and known variance

$$X_1, \ldots, X_n \sim \text{iid Normal}(M, \sigma^2)$$

- Prior $M \sim \text{Normal}(\mu_0, \sigma_0^2)$, continuous distribution
  - $f_M(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})$
- Samples: $x_1, \ldots, x_n$, Sample mean: $\overline{x} = (x_1 + \cdots + x_n)/n$
- Posterior: $M|(X_1 = x_1, \ldots, X_n = x_n)$ is continuous
  - Posterior density $\propto f(X_1 = x_1, \ldots, X_n = x_n|M = \mu) f_M(\mu)$
  - Posterior density $\propto \exp(-\frac{(x_1-\mu)^2+\cdots+(x_n-\mu)^2}{2\sigma^2}) \exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})$
- Posterior density: Normal
  - Posterior mean $= \overline{x}\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2} + \mu_0\frac{\sigma^2}{n\sigma_0^2+\sigma^2}$

$$\hat{\mu} = \frac{X_1 + \cdots + X_n}{n} \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} + \mu_0 \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}$$

Oberservations for Normal prior

- Prior: $\text{Normal}(\mu_0, \sigma_0^2)$
  - How to pick μ_0 and σ_0?
- Estimate is combination of data and prior
  - Prior is "updated" using data to get posterior
- If n is very large, $\hat{\mu} \to$ sample mean
  - Data dominates the estimate
  - Prior plays no significant role
- If n is small, prior contributes significantly to the estimate
  - Prior needs to have some justification when n is small
- If variance of prior is large compared to variance of samples, prior tends to be flat or uninformative
  - Choice of variance of prior is important

Section 3

Problems: Finding estimators

Problem 1

Suppose X is a discrete random variable taking values $\{0, 1, 2, 3\}$ with respective probabilities $\{2\theta/3, \theta/3, 2(1-\theta)/3, (1-\theta)/3\}$, where $0 \leq \theta \leq 1$ is a parameter. Consider the estimation of θ from samples $2, 2, 0, 3, 1, 3, 2, 1, 2, 3$.

$\bar{x} = 1.9$

- Find the method of moments and maximum likelihood estimates.
- Using a Uniform$[0, 1]$ prior, find the posterior distribution and mean.

Problem 2

Consider n iid samples from a Geometric(p) distribution.

- Find the method of moments estimate.
- Find the MLE.
- Using a Uniform$[0, 1]$ prior, find the posterior distribution and mean.

Problem 3

Consider n iid samples from a Poisson(λ) distribution.

- Find the method of moments estimate.
- Find the MLE.
- Using a Gamma$[\alpha, \beta]$ prior, find the posterior distribution and mean.

Section 4

Problems: Fitting distributions

Problem 1

Fit a Poisson distribution to the following frequency data on number of vehicles (n) making a right turn at an intersection in a 3-minute interval. Find an approximate 95% confidence interval for the sample mean using a normal approximation for the sampling distribution.

n	Frequency	n	Frequency
0	14	7	14
1	30	8	10
2	36	9	6
3	68	10	4
4	43	11	1
5	43	12	1
6	30	13+	0

$\hat{\lambda} = 3.893$

X = Number of Vehicles making a right turn in a 3-minute interval

$X \sim \text{Poisson}(\lambda)$

Problem 2

Fit a Geometric distribution to the following frequency data on number of hops (n) between flights of birds. Find an approximate 95% confidence interval.

n	Frequency	n	Frequency
1	48	7	4
2	31	8	2
3	20	9	1
4	9	10	1
5	6	11	2
6	5	12	1

Problem 3

Data from a genetic experiment and expected distribution in terms of an unknown parameter θ are given in the following table.

Type	Frequency	Theory
1	1997	$0.25(2+\theta)$
2	906	$0.25(1-\theta)$
3	904	$0.25(1-\theta)$
4	32	0.25θ

pmf for X

$0<\theta<1$

Section 5

Problems: Model and estimation

Problem 1

To find the size of an animal population, 100 animals are captured and tagged. Some time later, another 50 animals are captured, and 20 of them were found to be tagged. How will you estimate the population size? What are your assumptions?

Capture-recapture

Problem 2

In a new machine, suppose that, out of 10 produced items, no item was found to be defective. How will you estimate the fraction of defective items produced by the new machine? From data collected from other similar machines, the average of the fraction of defective items was found to be 10%, and the actual fraction was between 5% and 15% in 95% of the cases.

Hypothesis testing

Section 1

Introduction

What is hypothesis testing?

Motivating example: Is a coin authentic or counterfeit?

An authentic coin is known to have $P(H) = 0.5$ when tossed, while a counterfeit coin has $P(H) = 0.6$. Suppose you have a coin that could be authentic or counterfeit. You may toss the coin multiple times and observe the results. How will you test whether the coin is authentic or counterfeit?

Hypothesis testing

- Using samples, decide between a *null hypothesis* denoted H_0 and an *alternative hypothesis* denoted H_A
  - Counterfeit coin example: H_0: $P(H) = 0.5$ and H_A: $P(H) = 0.6$
- One of the most important statistical analysis methods with a wide range of applications

Accepting or Rejecting the Null Hypothesis

Example: Is a coin authentic or counterfeit?

- Suppose we toss the coin 3 times
  - Possible outcomes are $HHH, HHT, \ldots, TTT$
  - For some outcomes, we will accept H_0 and the others, we will reject H_0
  - Let A be the set of all outcomes for which we accept H_0
- Every *acceptance* subset A corresponds to a *test*

Acceptance set and test

$X_1, \ldots, X_n \sim$ iid X, H_0: null hypothesis, H_A: alternative

- Suppose $X \in \mathcal{X}$. Then, the samples $X_1, \ldots, X_n \in \mathcal{X}^n$
- Subset $A \subseteq \mathcal{X}^n \leftrightarrow$ a hypothesis test

If $X_1, \ldots, X_n \in A$, we accept H_0; otherwise, we reject H_0

Metrics for hypothesis testing

Example: Is a coin authentic or counterfeit?

H_0: $P(H) = 0.5$ and H_A: $P(H) = 0.6$

- Suppose we toss the coin 3 times: 8 outcomes
  - $2^8 = 256$ subsets $\leftrightarrow$ 256 tests
- How to define a *good* acceptance set or a test?

Size and power of a test

- Metric 1: *Significance level* (also called *size*) of a test, denoted α
  - Type I error: Reject H_0 when H_0 is true
  - $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{is true})$
- Metric 2: *Power* of a test, $1 - \beta$
  - Type II error: Accept H_0 when H_A is true
  - $\beta = P(\text{Type II error}) = P(\text{Accept } H_0 | H_A \text{is true})$
  - Power $= 1 - \beta = P(\text{Reject } H_0 | H_A \text{is true})$

Counterfeit coin: Computing α, β

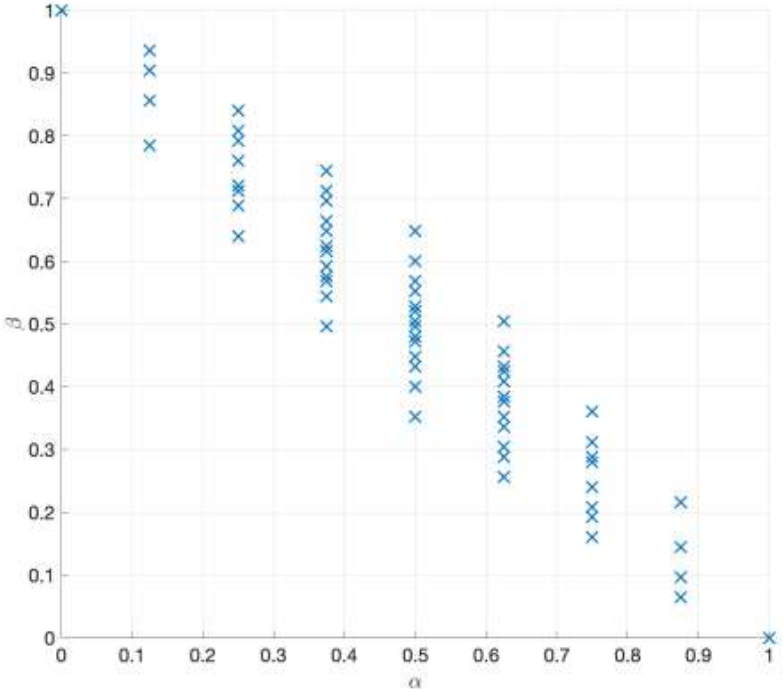
H_0: $P(H) = 0.5$ and H_A: $P(H) = 0.6$

Toss 3 times. $\mathcal{X}^3 = \{HHH, HHT, \ldots, TTT\}$

- $A = \phi$
  - Always reject H_0
  - $\alpha = 1$, $\beta = 0$
- $A = \mathcal{X}^3$
  - Always accept H_0
  - $\alpha = 0$, $\beta = 1$
- $A = \{HHT, HTH, HTT, THH, THT, TTH\}$
  - $\alpha = P(A^c|P(H) = 0.5) = 2/8 = 0.25$
  - $\beta = P(A|P(H) = 0.6) = 3(0.4)^2(0.6) + 3(0.4)(0.6)^2 = 0.72$
- $A = \{TTT, TTH, THT, HTT\}$
  - $\alpha = 4/8 = 0.5$
  - $\beta = 0.4^3 + 3(0.4)^2(0.6) = 0.352$

There is a tradeoff between α and β

Counterfeit coin: α, β for all 256 tests



What if we toss 100 times? What about other distributions?

Neyman-Pearson paradigm of hypothesis testing

$$X_1, \ldots, X_n \sim \text{iid } X$$

- H_0: null hypothesis on distribution of X, H_A: alternative hypothesis
- Test: defined by an acceptance set A
  - If samples fall in A, accept H_0; otherwise, reject H_0
- Two errors
  - Type I error: Reject H_0 when H_0 is true
  - Type II error: Accept H_0 when H_A is true
- Two metrics
  - *Significance level*, α
    - $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0)$
  - *Power* of a test, $1 - \beta$
    - $\beta = P(\text{Type II error}) = P(\text{Accept } H_0 | H_A)$

Section 2

Problems

Problem 1

Consider 100 tosses of a coin, which could be either authentic with probabiltiy of heads equal to 0.5, or counterfeit with probability of heads 0.6. Suppose T is the number of heads seen. Consider a test that rejects H_0 if $T > c$ for some constant c. What is the significance level of the test? What is the power of the test?

Problem 2

Consider one sample $X \sim \text{Normal}(\mu, 1)$. Let the null and alternative hypothesis be $H_0 : \mu = -1$ and $H_A : \mu = 1$. Consider a test that rejects H_0 if $X > c$ for some constant c. What is the significance level of the test? What is the power of the test?

Problem 3

Consider one sample $X \sim \text{Binomial}(100, p)$. Let the null and alternative hypothesis be $H_0 : p = 0.5$ and $H_A : p \neq 0.5$. Consider a test that rejects H_0 if $|X - 50| > 10$. What is the significance level of the test? What is the power of the test as a function of p? Use the normal approximation.

Problem 4

Consider 100 samples $X_1, \ldots, X_{100} \sim$ iid Normal$(\mu, 1)$. Let the null and alternative hypothesis be $H_0 : \mu = -1$ and $H_A : \mu = 1$. Suppose $T = (X_1 + \cdots + X_{100})/100$. Consider a test that rejects H_0 if $T > c$ for some constant c. What is the significance level of the test? What is the power of the test?

Section 3

Types of hypothesis testing

Simple hypothesis

Definition (Simple hypothesis)

A hypothesis that completely specifies the distribution of the samples is called a simple hypothesis.

Examples of simple hypothesis

- Coin toss
  - P(Heads) $= 0.5$, P(Heads) $= 0.9$ etc.
- Normal$(\mu, 3)$ samples
  - $\mu = 1$, $\mu = -1$ etc.

Simple null vs simple alternative

- Very well understood, best approach is known
- Rarely occurs

Composite hypothesis

Definition (Composite hypothesis)

A hypothesis that does not completely specify the distribution of the samples is called a composite hypothesis.

Examples

- Coin toss
  - Null: P(Heads) $= 0.5$ (coin is fair), simple
  - Alternative: P(Heads) $\neq 0.5$ (coin is unfair), composite
- Normal$(\mu, 3)$ samples
  - Null: $\mu = 0$ (some effect is not present), simple
  - Alternative: $\mu > 1$ (effect is present), composite

Simple/composite null vs composite alternative

- Well studied, but multiple approaches are possible
- Most common

Standard tests: One sample

$$X_1, \ldots, X_n \sim iid X, E[X] = \mu, \text{Var}(X) = \sigma^2$$

- Testing for mean, null $H_0 : \mu = c$
  - Alternative
    - ⋆ Right tail test, $H_A : \mu > c$
    - ⋆ Left tail test, $H_A : \mu < c$
    - ⋆ Two tail test, $H_A : \mu \neq c$
  - Two cases: known or unknown variance
- Testing for variance
  - Null $H_0 : \sigma = c$
  - Alternative $H_A : \sigma > c$

Standard tests: Two samples

$$X_1, \ldots, X_{n_1} \sim iid X, E[X] = \mu_1, \text{Var}(X) = \sigma_1^2$$

$$Y_1, \ldots, Y_{n_2} \sim iid Y, E[X] = \mu_2, \text{Var}(X) = \sigma_2^2$$

- Testing to compare means
  - Null $H_0 : \mu_1 = \mu_2$
  - Alternative $H_A : \mu_1 \neq \mu_2$
- Testing to compare variances
  - Null $H_0 : \sigma_1 = \sigma_2$
  - Alternative $H_A : \sigma_1 \neq \sigma_2$

Goodness of fit testing

Samples: $X_1, \ldots, X_n$

Problem: Do the samples follow a certain distribution?

Examples

- Integer samples $X_i \in \{0, 1, 2, \ldots\}$. Is the distribution Poisson?
- Continuous positive samples $X_i \in [0, \infty)$. Is the distribution Gamma?
- Continuous samples $X_i \in (-\infty, \infty)$. Is the distribution normal?
- Multinomial $X_i \in \{1, 2, \ldots, M\}$. Is the distribution $\{f_1(\theta), \ldots, f_M(\theta)\}$?

Section 4

Answering questions using data

Questions

- In most cases, useful questions can be posed as the testing of a hypothesis. Several classes of testing problems arise.
- Breaking down the question so that it becomes a hypothesis test is an important *design* step

Examples

- A person claims magical powers in being able to predict something. How will you design a statistical hypothesis test?
- A company claims a new treatment method for a disease. How will you test for the effectiveness of the treatment?
- Data of accidental deaths in a country: Is there a seasonal or monthly pattern in this data?
- Data on hiring by an organization: Is there any gender or geographical bias in the hiring?

Example 1: Magical powers

Suppose a person claims magical powers to predict the throw of a die. Here is one possible way to post it as a hypothesis testing problem.

- Throw the die n times and record the predictions. Let T be the number of correct predictions.
- Null $H_0 : T \sim \text{Binomial}(n, 1/6)$
- Alternative $H_A : T \sim$ any other distribution

In the above test, we need to measure or quantify the *confidence* of our conclusion and justify its *statistical significance*. The number of trials n will be an important factor to decide.

Example 2: New medical treatment

A company claims a new drug is effective in reducing heart attacks in a certain segment of the population. Here is a common way in which drugs are tested.

- n volunteers are chosen randomly for the trial from the population segment. About $n/2$ of them (Group I) are chosen randomly and given the drug, and the remaining (Group II) are given a placebo. Volunteers are not told what they got.
- Over a time period, the volunteers are observed for heart attacks. Suppose the fraction of volunteers who got a heart attack in Group I is f_1, and the same fraction in Group II is f_2
- Null $H_0 : f_1 \approx f_2$
- Alternative $H_A : \dfrac{f_2 - f_1}{f_2} = c$

It is important to find c, and to quantify *confidence* and *statistical significance*. Once again, n will be an important factor.

Example 3: Pattern in accidental deaths

The number of accidental deaths in a country are tabulated every month over a year. Here is one way to test if there is a constant number of deaths per day, i.e. a constant rate.

- Estimate the overall rate, $\theta =$ Total deaths / Total number of days
- Estimated monthly deaths: $\{31\theta, 28\theta, 31\theta, \ldots, 31\theta\}$
- Null H_0 : Estimated deaths fits the observed deaths
- Alternative H_A: Estimated and observed deaths do not fit for some months

Assessing goodness of fit is an important ingredient here. We need to quantify the *confidence* in the fit.

Example 4: Gender bias in hiring

Consider the following cross-tabulation of hires made by a company.

	Female	Male	Total
Hired	6	12	18
Not hired	9	25	34
Total	15	37	52

Is there a gender bias in the hiring? Here is an approach.

- Pick 18 out of 52 uniformly at random, $T = M - F$
- Null H_0 : Distribution of T is as given above
- Alternative H_A : Any other distribution

Is 6 a reasonable value for T? How to quantify this?

Observations

- In all examples, the question seems to be reasonably posed in a statistical hypothesis testing framework
- In most cases, the null and/or alternative are composite
- In all cases, the *confidence* of the testing is very important
- How do you quantify *confidence*?
  - We use ideas from confidence interval of estimation
  - A notion called P-value is used to quantify confidence

Section 5

Standard testing methods: z-test
General methodology of testing

$$X_1, \ldots, X_n \sim \text{iid } X$$

- *Test statistic*, denoted T
  - Some function of the samples
  - Example: Sample mean $\overline{X}$, Sample variance S^2 etc
- Acceptance and rejection regions are specified through T
  - Examples
    - Reject H_0 if $T > c$ (right)
    - Reject H_0 if $T < -c$ (left)
    - Reject H_0 if $|T| > c$ (two-sided)
- Significance level α depends on c and the distribution of $T|H_0$
  - Right-sided: $\alpha = P(T > c|H_0)$ (similar for others)
  - Fix α and find c

Testing for mean (normal samples, known variance)

$$X_1, \ldots, X_n \sim \text{iid } N(\mu, 4^2)$$

- Test statistic $T = \overline{X} \triangleq \frac{1}{n}(X_1 + \cdots + X_n)$
- Null $H_0 : \mu = 0$, Alternative $H_A : \mu > 0$
- Test: Reject H_0 if $T > c$
- Different samplings, $n = 10$
  - [-6.9, 0.6, -0.6, -4.8, -1.9, -5.1, 7.5, 6.1, 0.5, 3.3], $T = -0.14$
  - [-1.8, -1.8, 4.1, 3.4, 1.9, 0.6, 1.7, -6.9, 0.3, -4.0], $T = -0.25$
  - [-5.8, 2.0, 2.5, 1.7, -2.8, 0.9, -0.4, 0.6, -8.5, -2.9], $T = -1.25$
  - [4.2, 14.2, 7.1, -5.1, -2.3, -3.9, -3.2, -0.9, -1.4, -6.4], $T = 0.23$
  - [1.0, 3.6, 5.9, -2.2, 2.3, 6.9, 1.7, 0.1, 6.3, 4.0], $T = 2.96$
  - [1.7, 3.9, -1.6, 3.8, 4.0, 1.9, -1.8, 10.3, 4.2, 4.6], $T = 3.10$
  - [9.4, 2.2, 13.8, 3.1, 6.3, 7.0, 5.8, 1.0, 7.6, 5.7], $T = 6.20$

Higher values of T give us more confidence in rejecting null

Testing for mean: Significance level

$$X_1, \ldots, X_{10} \sim \text{iid } N(\mu, 4^2)$$

- Significance level $\alpha = P(\overline{X} > c | \mu = 0)$
- Since $(\overline{X} | \mu = 0) \sim \text{Normal}(0, 4^2/10)$, we have

$$\alpha = P\left(\frac{\overline{X}}{4/\sqrt{10}} > \frac{c}{4/\sqrt{10}}\right) = 1 - F_Z(\sqrt{10}c/4)$$

c → critical value

c	0	1.62	2.08	2.94	3.26	3.91
α	0.5	0.1	0.05	0.01	0.005	0.001

z-test at significance level α: Reject H_0 if $T > c$, where c is as above.

Testing for mean: Results and P-value

c	0	1.62	2.08	2.94	3.26	3.91
α	0.5	0.1	0.05	0.01	0.005	0.001
$T = 0.23$	Rej	Acc	Acc	Acc	Acc	Acc
$T = 2.96$	Rej	Rej	Rej	Rej	Acc	Acc
$T = 6.20$	Rej	Rej	Rej	Rej	Rej	Rej

Definition (P-value)

Suppose the test statistic $T = t$ in one sampling. The lowest significance level α at which the null will be rejected for $T = t$ is said to be the P-value of the sampling.

- Finding P-value for $T = t$: Put $c = t$ in computation of α

T	-0.14	0.23	2.96	6.20
P-value	0.544	0.428	0.00964	4.755e-07

If P-value is low enough, we reject H_0

What to choose? Significance level or P-value

- Samples are given, and there is some hypothesis that needs to be tested
- Step 1: Decide on the null and alternative hypotheses H_0 and H_A
- Step 2: Decide on the test statistic T
- Step 3: "Philosophy" of testing
  - Choice 1: Pick a significance level first
    - ★ Probability of Type I error can be fixed in some applications. In those cases, significance level is easy to fix
    - ★ Historically, in many applications, 0.05 or 0.01 is accepted as a common significance level
    - ★ Find rejection region (find the *critical value* c and reject H_0 if $T > c$, for example)
  - Choice 2: Use P-value
    - ★ Report the P-value
    - ★ If P-value is *low enough*, choose to reject H_0; otherwise, accept H_A
    - ★ How low is low enough? Depends on applications and other information

Section 6

z-test problems

Problem 1

Suppose $X \sim \text{Normal}(\mu, 9)$. For $n = 16$ iid samples of X, the observed sample mean is 10.2. What conclusion would a z-test reach if the null hypothesis assumes $\mu = 9.5$ (against an alternative hypothesis $\mu > 9.5$) at a signifiance level of $\alpha = 0.05$? What if the null hypothesis assumes $\mu = 8.5$ (against an alternative hypothesis $\mu > 8.5$)?

Problem 2

Suppose an app is desired to make an accurate identification of faces in photographs more than 90% of the time in the long run. For a random sample of 500 such photos, the app makes the correct identification 462 times - a 92.4% success rate. What does a z-test say about a null hypothesis that the app is only 90% accurate (compared to an alternative hypothesis that the app is more than 90% accurate with a significance level of $\alpha = 0.05$)?

Samples: $X_1, \cdots, X_{500} \overset{iid}{\sim} \text{Bernoulli}(p)$ $\quad X_i = \begin{cases} 1 & \text{if fp correctly identified} \\ 0 & \text{else} \end{cases}$

Problem 3

Suppose $X \sim \text{Normal}(\mu, 36)$. For $n = 25$ iid samples of X, the observed sample mean is 6.2. What conclusion would a z-test reach if the null hypothesis assumes $\mu = 4$ (against an alternative hypothesis $\mu \neq 4$) at a signifiance level of $\alpha = 0.05$? What if the null hypothesis assumes $\mu = 8$ (against an alternative hypothesis $\mu < 8$)?

Section 7

More problems on z-test

Problem 1a (with binomial distribution)

Vaccine hesitancy (percentage of people who are unwilling to vaccinate) in a town has been reported to be 20%. To test whether the fraction is 20%, you call a randomly selected group of 10 people and find out that 3 of them are vaccine hesitant. What is the null hypothesis? Will you accept or reject null at a significance level of $\alpha = 0.05$ against an alternative that the fraction is above 20%? What is power against an alternative that the fraction is 30%?

Problem 1b (with normal approximation)

Vaccine hesitancy (percentage of people who are unwilling to vaccinate) in a town has been reported to be 20%. To test whether the fraction is 20%, you call a randomly selected group of 100 people and find out that 28 of them are vaccine hesitant. What is the null hypothesis? Will you accept or reject null at a significance level of $\alpha = 0.05$ against an alternative that the fraction is above 20%? What is power against an alternative that the fraction is 30%?

$T = \text{Binomial}(100, p) \approx \text{Normal}(100p, 100p(1-p))$

Problem 2a

The current-carrying capacity of a resistor manufactured at your company is supposed to be 3.0 A (A stands for Amperes). Because of a recent change in the manufacturing process, you suspect that the current-carrying capacity might actually be lesser than 3.0 A. You decide to test by measuring current-carrying capacities of 10 resistors with a test measurement has a standard deviation of 0.05 A. If the sample mean of the test measurements $T_{10} < 2.95$, you will conclude that the manufacturing process is faulty.

1. What is the null hypothesis? What is the alternative hypothesis? What are the samples?
2. What is the significance level α of the test?
3. If the current-carrying capacity falls to 2.9 A, there could be serious safety issues. Against the alternative hypothesis of 2.9 A, what is the power $(1 - \beta)$ of the test?

Samples: $X_1, X_2, \ldots, X_{10} \sim N(\mu, 0.05^2)$ X_i = measured capacity of i-th test resistor

Problem 2b

In the same problem, suppose you can test n resistors. If the sample mean $T_n < c$, you conclude that the manufacturing process is faulty. You need to determine suitable values for n and c under the following conditions:

1. Significance level or probability of Type I error, $\alpha \leq 10^{-6}$
2. Probability of Type II error, $\beta \leq 10^{-12}$ (against an alternative of 2.9 A)

Problem 3

The average CGPA of students in a college is reported to be 8.0 with a standard deviation of 1. You suspect that the average may be lower, possibly 7.5, and decide to sample students to find their CGPA. What sample size do you need for a test at a significance level of 0.05 and power of 0.95? How will the sample size change if you suspect the CGPA to be 7.0?