

Math 4261: Mathematical Statistics I

Lecture 5: Sampling distributions, MGFs, and modes of convergence

Ref: Blitzstein–Hwang 6.4; HMC 5.1–5.2

Many ideas and figures credited to Zhou Fan

Fall 2026

Outline

Recap

Motivation: Sampling distributions

Moment generating functions

Approximations with tail estimates

Modes of convergence

Outline

Recap

Motivation: Sampling distributions

Moment generating functions

Approximations with tail estimates

Modes of convergence

Recap: Multivariate Gaussian

Theorem

Suppose $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in $\mathbb{R}^p$. Let $\mathbf{a} \in \mathbb{R}^q$ and $\mathbf{B} \in \mathbb{R}^{q \times p}$ be fixed. Then

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim \mathcal{N}(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top).$$

We record a useful special case: when $\mathbf{b} \in \mathbb{R}^p$,

$$\mathbf{b}^\top \mathbf{X} \sim \mathcal{N}(\mathbf{b}^\top \boldsymbol{\mu}, \mathbf{b}^\top \boldsymbol{\Sigma} \mathbf{b}).$$

- ▶ Every linear function of a multivariate Gaussian remains Gaussian.
- ▶ For a multivariate Gaussian, **uncorrelated** $\iff$ **independent**.

Recap: Best estimates

Let X be observed data and Y the quantity of interest to predict. We choose an estimator $\delta(X)$ to minimize

$$\text{MSE}(\delta) = \mathbb{E}[(Y - \delta(X))^2].$$

The **best estimator** is

$$\delta_{\text{opt}}(X) = \mathbb{E}[Y \mid X].$$

If we restrict to **constant estimators**, the best estimator is the mean:
 $c_{\text{opt}} = \mathbb{E}[Y]$.

The **best linear estimator** is

$$\delta_{\text{lin}}(X) = \mathbb{E}[Y] + \frac{\text{Cov}[X, Y]}{\text{Var } X}(X - \mathbb{E}X).$$

Outline

Recap

Motivation: Sampling distributions

Moment generating functions

Approximations with tail estimates

Modes of convergence

Sampling distributions of statistics

For data $X_1, \dots, X_n$, a **statistic** $T(X_1, \dots, X_n)$ is any function of the data. For instance

$$\text{(sample mean)} \quad \bar{X} = \frac{X_1 + \dots + X_n}{n},$$

$$\text{(sample variance)} \quad S^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1},$$

$$\text{(sample range)} \quad R = \max(X_1, \dots, X_n) - \min(X_1, \dots, X_n).$$

If the data $X_1, \dots, X_n$ are random, then this randomness induces a **sampling distribution** for the statistic.

If we understand the randomness of $X_1, \dots, X_n$, how can we understand the sampling distribution of $T(X_1, \dots, X_n)$?

Example: Gaussian sample

Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. What is the distribution of the sample mean

$$\overline{X} = \frac{X_1 + \dots + X_n}{n}?$$

We may compute

$$\begin{aligned}\mathbb{E}[\overline{X}] &= \frac{\mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]}{n} = \mu, \\ \text{Var}(\overline{X}) &= \frac{\text{Var}(X_1) + \dots + \text{Var}(X_n)}{n^2} = \frac{\sigma^2}{n}.\end{aligned}$$

From last lecture, any linear combination of independent normal variables has a normal distribution. So

$$\boxed{\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)}.$$

Example: Political poll

In the first lecture we considered estimating the support for Harris in the 2024 election. Suppose we drew a simple random sample (**with replacement**) $X_1, \dots, X_n$ from the population of size N with unknown support θ . Thus

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(\theta).$$

Thus the **sampling distribution**, the distribution of $\hat{\theta}$, is a **scaled Binomial** random variable:

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n) = \frac{X_1 + \dots + X_n}{n} \sim \frac{1}{n} \text{Bin}(n, \theta).$$

Overview

Today we build tools and learn frameworks to understand **sampling distribution** of a statistic:

1. **Exact distribution.** Tool: moment generating functions.
2. **Finite sample control.** When an exact probability is difficult to compute, use a tractable bound.
Application: Gaussian tail estimates.
3. **Large sample approximation.** As the sample size grows, asymptotic tools become useful.
Today: modes of convergence. Next time: LLN and CLT.

Outline

Recap

Motivation: Sampling distributions

Moment generating functions

Approximations with tail estimates

Modes of convergence

Moment generating functions

For any r.v. X , its **moment generating function** is

$$M_X(t) = \mathbb{E}\left[e^{tX}\right],$$

for those $t \in \mathbb{R}$ where this expectation is finite.

- ▶ The MGF carries a huge amount of information about a distribution.
- ▶ If $M_X(t) < \infty$ in an interval around 0, plugging in the Taylor series for exponential gives

$$M_X(t) = \mathbb{E}e^{tX} = 1 + t\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \frac{t^3}{3!}\mathbb{E}[X^3] + \dots$$

$M_X(t)$ is a power series in t , with coefficients determined by the moments.

Usefulness of MGF

The MGF *generates* all the moments of X :

Theorem

If $M_X(t) < \infty$ in an interval around $t = 0$, then

$$\mathbb{E}[X^k] = \left. \frac{d^k}{dt^k} M_X(t) \right|_{t=0} = M_X^{(k)}(0).$$

The MGF of a random variable **determines its distribution**:

Theorem

Let X and Y be two r.v.'s such that $M_X(t) = M_Y(t) < \infty$ in an interval around $t = 0$. Then they have the same distribution.

Examples:

- ▶ If X is **Bernoulli** distributed, then $\mathbb{E}[X] = p$ determines the distribution.
- ▶ If X is **Gaussian** distributed, then $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ determine the distribution (variance is a function of the first two moments).

Example

Let $X \sim \text{Bern}(p)$. Find $M_X(t)$.

Let $q = 1-p$

$$M_X(t) = \mathbb{E}[e^{tX}]$$

$$\stackrel{\text{thus}}{=} e^{t \cdot 1} \overbrace{p[X=1]}^p + e^{t \cdot 0} \overbrace{p[X=0]}^q$$

$$= pe^t + q$$

Sanity check: $X = X^2 = X^3 = \dots = X^k = \dots$, so $\mathbb{E}X = p = \mathbb{E}X^k$ for all k .

$$M_X(t) = 1 + t \overbrace{\mathbb{E}X}^p + \frac{t^2}{2} \overbrace{\mathbb{E}X^2}^p + \frac{t^3}{3!} \overbrace{\mathbb{E}X^3}^p + \dots$$

$$= 1 + p \left(t + \frac{t^2}{2} + \frac{t^3}{3!} + \dots \right)$$

$$= pe^t + qp$$

MGF and independence

If X and Y are independent, then the MGF of $X + Y$ is the product of the individual MGFs:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Example: Binomial

Let $X \sim \text{Bin}(n, p)$. Since X is a sum of n independent Bernoulli random variables $X = X_1 + \cdots + X_n$,

$$M_X(t) = M_{X_1}(t) \cdots M_{X_n}(t) = (1 - p + pe^t)^n.$$

Example

Let $Z \sim \mathcal{N}(0, 1)$. Find $M_Z(t)$.

$$M_X(t) = \mathbb{E} e^{tX} \stackrel{\text{LOTUS}}{=} \int_{\mathbb{R}} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp \left\{ -\frac{1}{2}(x-t)^2 + \frac{t^2}{2} \right\} dx$$

$$= e^{t^2/2} \int_{\mathbb{R}} \underbrace{\frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}(x-t)^2 \right\}}_{\text{PDF of } \mathcal{N}(t, 1)} dx$$

PDF of $\mathcal{N}(t, 1)$

$= 1$

$$= e^{t^2/2}$$

Example

Let $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

Recall $Z_1^2 + \dots + Z_n^2 \sim \chi_n^2$. We now show it.

Equivalently, χ_n^2 is $\text{Gamma}(n/2, 1/2)$.

Fact: Let $Y \sim \text{Gamma}(\alpha, \beta)$. Then $M_Y(t) = \begin{cases} \infty & \text{if } t \geq \beta \\ \frac{1}{(1 - \frac{t}{\beta})^\alpha} & \text{if } t < \beta \end{cases}$

Consider $n=1$: χ_1^2

$$M_{\chi_1^2}(t) = \mathbb{E} e^{t\chi_1^2} = \int_{\mathbb{R}} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{(t - \frac{1}{2})x^2} dx = \infty \text{ if } t \geq \frac{1}{2}$$

$$\frac{1}{\sqrt{1-2t}} \sqrt{\frac{1-2t}{2\pi}} e^{-\frac{1}{2}(1-2t)x^2}$$

PDF of $\mathcal{N}(0, \frac{1}{1-2t})$

$$= \begin{cases} \infty & t \geq \frac{1}{2} \\ \frac{1}{(1-2t)^{1/2}} & t < \frac{1}{2} \end{cases}$$

χ_1^2
||
MGF of $\text{Gamma}(\frac{1}{2}, \frac{1}{2})$.

Example

Hence $X_1^2 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$.

General n :

$$M_{X_1^2 + \dots + X_n^2}(t) = \mathbb{E} e^{t(X_1^2 + \dots + X_n^2)}$$

$$= \underbrace{\mathbb{E} [e^{tX_1^2}]}_{\text{M.F.F. of } \text{Gamma}(\frac{1}{2}, \frac{1}{2})} \times \dots \times \mathbb{E} [e^{tX_n^2}]$$

$$= \begin{cases} \infty & t \geq \frac{1}{2} \\ \frac{1}{(1-2t)^{1/2}} & t < \frac{1}{2} \end{cases}$$

$$= \begin{cases} \infty & t \geq \frac{1}{2} \\ \frac{1}{(1-2t)^{n/2}} & t < \frac{1}{2} \end{cases}$$

M.F.F. of $\text{Gamma}(\frac{n}{2}, \frac{1}{2})$

Hence $X_1^2 + \dots + X_n^2 \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$.

Outline

Recap

Motivation: Sampling distributions

Moment generating functions

Approximations with tail estimates

Modes of convergence

The need for approximation

For many (seemingly simple) statistics, the sampling distribution is **difficult** to describe exactly. For example:

1. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Unif}(-1, 1)$. What is the distribution of the sample mean $\bar{X}$?
2. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(1/2)$, the outcomes of n tosses of a fair coin. What is the distribution of

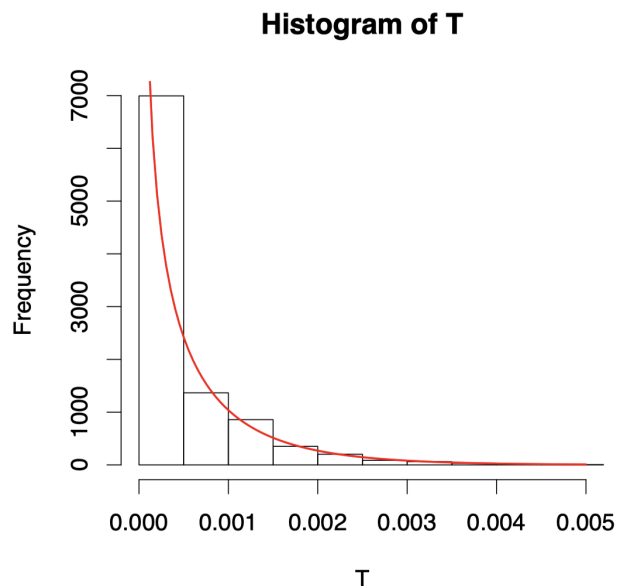
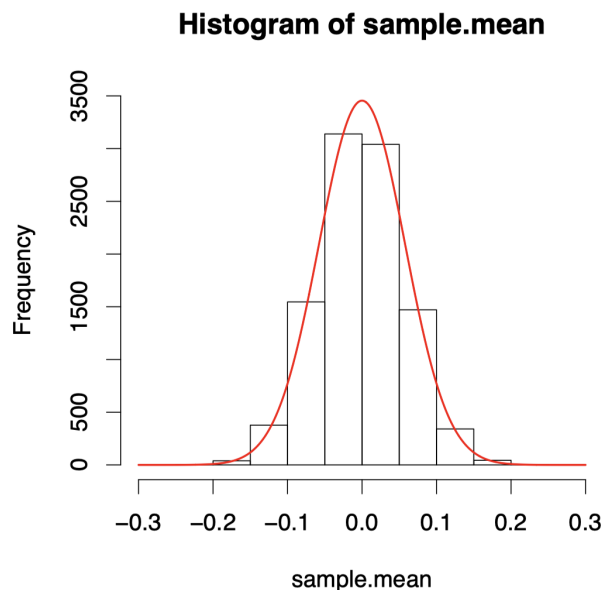
$$T = \left(\bar{X} - \frac{1}{2} \right)^2 ?$$

If we compute T from n observed coin tosses and this is too large compared with typical values from this distribution, then we have evidence that the coin may not be fair. (This is the idea of hypothesis testing, which we'll discuss later in the course.)

For distributions that are difficult to study exactly, we can try to study them via computer simulation or large-sample approximation.

Large sample approximations

Often, a good approximation to the distribution of $T = T(X_1, \dots, X_n)$ emerges when the sample size n is large. These results are called **asymptotic** or **large sample approximations**.



The large sample approximations in this course will be based on two fundamental results from probability theory: the **LLN** and **CLT**.

More deviation inequalities

Large sample approximations describe what happens as n grows to infinity. What can we say when n is finite?

A **finite sample bound** controls an error probability for a fixed n .
If $\hat{\theta}_n$ estimates θ , the probability of a large error is the **tail probability**

$$\mathbb{P}\left[|\hat{\theta}_n - \theta| \geq \epsilon\right].$$

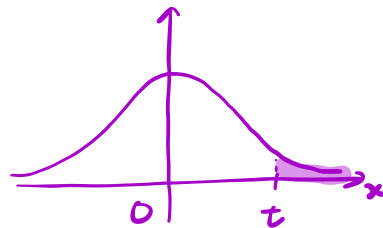
Often, an exact value is difficult to compute, and approximation is sufficient.

It turns out that **MGFs** are intimately related to tail probabilities.
We illustrate this with the Gaussian case.

Example: Gaussian tail bound

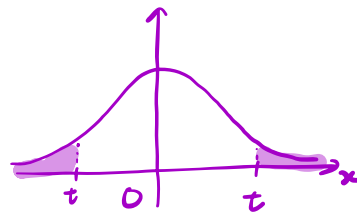
Let $Z \sim \mathcal{N}(0, 1)$. Then for any $t \geq 0$,

$$\mathbb{P}[Z \geq t] \leq \exp\left(-\frac{t^2}{2}\right).$$



Consequently, by **symmetry**,

$$\mathbb{P}[|Z| \geq t] \leq 2 \exp\left(-\frac{t^2}{2}\right).$$



$$\begin{aligned} \textcircled{1} \mathbb{P}[X \geq t] &= \mathbb{P}[\beta X \geq \beta t] && \text{(for } \beta > 0 \text{ TBD)} \\ &= \mathbb{P}[\exp(\beta X) \geq \exp(\beta t)] \stackrel{\text{Markov}}{\leq} \frac{\mathbb{E} \exp(\beta X)}{\exp(\beta t)} \end{aligned}$$

Example: Gaussian tail bound

$$\mathbb{P}[Z \geq t] \leq \exp\left(-\frac{t^2}{2}\right).$$

Aside: Why do this?

say for $Y \geq 0$

$$\text{Markov: } \mathbb{P}[Y \geq a] \leq \frac{\mathbb{E}Y}{a}, \quad \text{first moment}$$

$$\text{Chebyshev: } \mathbb{P}[Y \geq a] \leq \frac{\text{Var } Y}{(a - \mathbb{E}Y)^2} = \frac{\mathbb{E}[Y^2] - (\mathbb{E}Y)^2}{(a - \mathbb{E}Y)^2} \quad \text{first and second moments}$$

The exp. function encapsulates information about all moments:

$$\exp Y = 1 + Y + \frac{Y^2}{2!} + \frac{Y^3}{3!} + \frac{Y^4}{4!} + \dots$$

Example: Gaussian tail bound ① $P[X \geq t] \leq \frac{\mathbb{E} \exp(\beta X)}{\exp(\beta t)}$ for $\beta \geq 0$.

$$\mathbb{P}[Z \geq t] \leq \exp\left(-\frac{t^2}{2}\right).$$

Gaussian
MGF

$$\textcircled{2} \mathbb{E} \exp(\beta X) = \exp\left(\frac{\beta^2}{2}\right)$$

$$\textcircled{3} P[X \geq t] \leq \exp\left(\frac{\beta^2}{2} - \beta t\right) \text{ holds } \forall \beta \geq 0$$

$$\leq \min_{\beta \geq 0} \exp\left(\frac{\beta^2}{2} - \beta t\right) \quad \frac{d}{d\beta}\left(\frac{\beta^2}{2} - \beta t\right) = \beta - t \stackrel{\text{set}}{=} 0$$

$$= \exp \min_{\beta \geq 0} \left(\frac{\beta^2}{2} - \beta t\right)$$

$$= \exp\left(\frac{t^2}{2} - t^2\right) = \exp\left(-\frac{t^2}{2}\right).$$



Example: Gaussian finite sample guarantee

Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where you know σ^2 , but don't know μ . How many **samples** n do you need to guarantee that you can estimate μ within ϵ accuracy, with probability at least 95%?

Use $\bar{X}$ to estimate μ

$$\text{Let } Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

$$P[|\bar{X} - \mu| > \epsilon] = P\left[|Z| \geq \frac{\sqrt{n}\epsilon}{\sigma}\right] \overset{\text{Gaussian tail bound}}{\leq} 2 \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) \overset{\text{Goal}}{\leq} 0.05$$

Solving :

$$n \geq \frac{2\sigma^2}{\epsilon^2} \log \frac{1}{0.05}$$

General mechanism: Chernoff bound

Generally, when we control the MGF, we control the **tails of the distribution**. Abstracting what we just did:

For any $s > 0$, the function $x \mapsto e^{sx}$ is increasing and nonnegative. Therefore by Markov's inequality

$$\begin{aligned}\mathbb{P}[X \geq a] &= \mathbb{P}[e^{sX} \geq e^{sa}] \\ &\leq \frac{\mathbb{E}[e^{sX}]}{e^{sa}} = e^{-sa} M_X(s).\end{aligned}$$

This holds for **every** $s > 0$ for which $M_X(s) < \infty$, so we may choose the best value:

Chernoff bound

$$\mathbb{P}[X \geq a] \leq \exp \left\{ \inf_{s>0} \{-sa + \log M_X(s)\} \right\}.$$

Outline

Recap

Motivation: Sampling distributions

Moment generating functions

Approximations with tail estimates

~~Modes of convergence~~ *next time*