STA 131A: Introduction to Probability Theory

Lecture 21: Moment Generating Functions

Dogyoon Song

Spring 2026, UC Davis

Announcements

Midterm 2

- Solutions and scores are posted online
- Please review the solution carefully and identify topics to revisit
- You may review your graded exam in discussion section tomorrow
- Regrade requests, if any, should follow the instructions posted on Canvas

Homework 6 is posted

- This includes re-working Midterm 2 problems and reflecting on what you might have missed and what you have reviewed
- Start early and post questions on Piazza as needed

Office hours today

- 2:30–3:30 PM at MSB 4220

Agenda

Last time:

- Covariance and correlation
- Variance of sums
- Conditional variance

Today:

- **Variance decomposition:** finishing the story of conditional variance
  - Law of total variance
  - Estimation/regression interpretation
- **Moment generating functions:** packaging moments into one function
  - Definition and examples
  - Computing moments from MGFs
  - Identifying distributions from MGFs

Recap: Conditional variance

Definition

The **conditional variance** of X given $Y = y$ is

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}\left[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y\right].$$

Equivalently,

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}[X^2 \mid Y = y] - (\mathbb{E}[X \mid Y = y])^2.$$

- $\mathrm{Var}(X \mid Y = y)$ quantifies the remaining uncertainty in X after observing $Y = y$
- $\mathrm{Var}(X \mid Y)$ is a random variable that takes value $\mathrm{Var}(X \mid Y = y)$ when $Y = y$

Question: How does the "total uncertainty" $\mathrm{Var}(X)$ relate to the collection of conditional variances $\mathrm{Var}(X \mid Y = y)$?

Law of total variance

Theorem (Law of total variance)

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

- $\mathbb{E}[\mathrm{Var}(X \mid Y)]$: average uncertainty remaining after observing Y.
- $\mathrm{Var}(\mathbb{E}[X \mid Y])$: variability explained by how the conditional mean changes with Y.

Interpretation:

total variability = average within-condition variability + between-condition variability.

This is the variance analogue of the law of total expectation:

$$\mathbb{E}[X] = \mathbb{E}\big[\,\mathbb{E}[X \mid Y]\,\big].$$

Proof: Why the law of total variance holds

Theorem (Law of total variance)

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

Proof. Let $m(Y) = \mathbb{E}[X \mid Y]$. Then

$$\mathrm{Var}(X \mid Y) = \mathbb{E}[X^2 \mid Y] - m(Y)^2.$$

Taking expectations and using the law of total expectation,

$$\mathbb{E}[\mathrm{Var}(X \mid Y)] = \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\big] - \mathbb{E}[m(Y)^2] = \mathbb{E}[X^2] - \mathbb{E}[m(Y)^2]. \tag{1}$$

Also, since $\mathbb{E}[m(Y)] = \mathbb{E}[X]$ by total expectation,

$$\mathrm{Var}(\mathbb{E}[X \mid Y]) = \mathrm{Var}(m(Y)) = \mathbb{E}[m(Y)^2] - \big(\mathbb{E}[m(Y)]\big)^2 = \mathbb{E}[m(Y)^2] - (\mathbb{E}[X])^2. \tag{2}$$

Adding the two equations (1) and (2) gives

$$\mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathrm{Var}(X).$$

Example: Total variance in the two-group model

Example

Recall the two-group model:

$$Y \sim \text{Bernoulli}(1/2), \qquad X \mid Y = 0 \sim N(0,1), \qquad X \mid Y = 1 \sim N(2,4).$$

Thus,

$$\mathbb{E}[X \mid Y] = 2Y, \qquad \text{Var}(X \mid Y) = 1 + 3Y.$$

Since $Y \sim \text{Bernoulli}(1/2)$,

$$\mathbb{E}[\text{Var}(X \mid Y)] = \mathbb{E}[1 + 3Y] = 1 + 3 \cdot \frac{1}{2} = \frac{5}{2}.$$

Also,

$$\text{Var}(\mathbb{E}[X \mid Y]) = \text{Var}(2Y) = 4\text{Var}(Y) = 4 \cdot \frac{1}{4} = 1.$$

Therefore, by the law of total variance,

$$\text{Var}(X) = \frac{5}{2} + 1 = \frac{7}{2}.$$

Estimation viewpoint: Resolved vs. remaining variability

Inverse-problem setting: Suppose X is an unknown quantity of interest, and Y is an observed measurement/data. Before observing Y, uncertainty about X is measured by $\mathrm{Var}(X)$.

Estimation of X given Y: After observing Y, a natural prediction of X is the conditional mean

$$\widehat{X} = \mathbb{E}[X \mid Y].$$

Define the residual $R = X - \widehat{X}$. Then

$$\mathbb{E}[R \mid Y] = \mathbb{E}[X - \mathbb{E}[X \mid Y] \mid Y] = 0.$$

Thus, after using the information in Y, the residual has conditional mean zero.

Moreover, since $\mathbb{E}[R] = 0$, $\mathrm{Var}(R) = \mathbb{E}[R^2]$. Therefore,

$$\mathrm{Var}(\widehat{X}) = \mathrm{Var}(\mathbb{E}[X \mid Y]) \qquad \text{and} \qquad \mathrm{Var}(R) = \mathbb{E}[R^2] = \mathbb{E}\left[\mathrm{Var}(X \mid Y)\right].$$

Therefore, the law of total variance can be read as

$$\mathrm{Var}(X) = \underbrace{\mathrm{Var}(\widehat{X})}_{\text{uncertainty resolved/explained by } Y} + \underbrace{\mathrm{Var}(R)}_{\text{average posterior uncertainty left after } Y}.$$

Message: Observing Y changes the conditional mean $\mathbb{E}[X \mid Y]$. The variability of these conditional means is the explained part; the residual term is the average uncertainty left after observing Y.

Regression viewpoint: Explained vs. residual variability

Regression setting: Suppose now that

$$Y = \text{response}, \qquad X = \text{predictor}.$$

The conditional mean

$$m(X) = \mathbb{E}[Y \mid X]$$

is the regression function: it gives the mean response after seeing the predictor value.

The law of total variance becomes

$$\mathrm{Var}(Y) = \underbrace{\mathrm{Var}(m(X))}_{\text{variation explained by the predictor}} + \underbrace{\mathbb{E}[\mathrm{Var}(Y \mid X)]}_{\text{average within-predictor noise}}.$$

Interpretation:

- If X is informative, then $m(X)$ changes substantially with X, so $\mathrm{Var}(m(X))$ is large.
- If responses are tightly concentrated around $m(X)$, then $\mathbb{E}[\mathrm{Var}(Y \mid X)]$ is small.

This is the probability-theory version of the familiar idea behind regression:

$$\text{total variation} = \text{explained variation} + \text{residual variation}.$$

Pop-up quiz: Total variance

Suppose that

$$\mathrm{Var}(X \mid Y) = 2 \quad \text{for every value of } Y, \qquad \mathbb{E}[X \mid Y] = 3Y, \qquad \mathrm{Var}(Y) = 4.$$

Question: What is $\mathrm{Var}(X)$?

A) 2
B) 6
C) 14
D) 38

Answer: D.
By the law of total variance,

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]) = 2 + \mathrm{Var}(3Y) = 2 + 9 \cdot 4 = 38.$$

Follow-up: In

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]),$$

which term becomes small if observing Y makes X almost deterministic?

Motivation for moment generating functions

We have computed moments directly:

$$\mathbb{E}[X], \qquad \mathbb{E}[X^2], \qquad \mathrm{Var}(X).$$

But repeating direct integration or summation for each can be tedious.

A **moment generating function** packages all moments into a single function:

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

Why is this useful?

- Differentiating M_X gives moments
- MGFs can identify distributions
- They can make sums of independent random variables easier; we will use this next lecture

Moment generating function

Definition (Moment generating function)

The **moment generating function** (or **MGF**) of a random variable X is

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

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

- Discrete case:
$$M_X(t) = \sum_x e^{tx} p_X(x).$$
- Continuous case:
$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx.$$

The MGF is a function of the parameter t, not a single number.

- $M_X(0) = 1$ always.
- Some random variables do not have a finite MGF for some $t \in \mathbb{R}$.
- To be useful as an MGF, $M_X(t)$ should be finite on an open interval around 0.

Examples: Bernoulli and Poisson MGFs

Example (Bernoulli MGF)

Let $X \sim \text{Bernoulli}(p)$. Then

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p.$$

Thus,

$$M_X(t) = \mathbb{E}[e^{tX}] = (1 - p)e^{t \cdot 0} + pe^{t \cdot 1} = 1 - p + pe^t.$$

Example (Poisson MGF)

Let $X \sim \text{Poisson}(\lambda)$, so

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

Then

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} p_X(k) = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \frac{\lambda^k}{k!}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = \exp\big[\, \lambda(e^t - 1) \,\big].$$

Example: Exponential MGF

Example

Let $X \sim \text{Exponential}(\lambda)$, so

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

For $t < \lambda$,

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx \\ &= \lambda \int_0^\infty e^{-(\lambda - t)x}\, dx \\ &= \frac{\lambda}{\lambda - t}. \end{aligned}$$

For $t \geq \lambda$, the integral diverges. Thus the MGF for $\text{Exponential}(\lambda)$ exists only for $t < \lambda$.

Example: Standard normal MGF

Example (Standard normal MGF)

Let $Z \sim N(0,1)$, so

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \qquad z \in \mathbb{R}.$$

Then

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(z-t)^2}{2} + \frac{t^2}{2} \right] dz \\ &= \exp\left(\frac{t^2}{2}\right) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(z-t)^2}{2} \right] dz}_{=1, \text{ normal density with mean } t} \\ &= e^{t^2/2}. \end{aligned}$$

MGFs generate moments

From MGFs to moments

When the MGF exists in an interval around 0, differentiating under the expectation gives

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

Example: In particular,

$$M_X'(t) = \mathbb{E}[Xe^{tX}], \qquad M_X''(t) = \mathbb{E}[X^2 e^{tX}].$$

Setting $t = 0$,

$$M_X'(0) = \mathbb{E}[X], \qquad M_X''(0) = \mathbb{E}[X^2].$$

Therefore,

$$\mathrm{Var}(X) = M_X''(0) - (M_X'(0))^2.$$

Takeaway: With MGF, all moments can be computed via differentiation.

Example: Moments from Bernoulli MGFs

Example (Bernoulli)

For $X \sim \text{Bernoulli}(p)$,

$$M_X(t) = 1 - p + pe^t.$$

Thus,

$$M_X'(t) = pe^t, \qquad M_X''(t) = pe^t.$$

Therefore,

$$\mathbb{E}[X] = M_X'(0) = p$$

and

$$\begin{aligned} \text{Var}(X) &= M_X''(0) - \left(M_X'(0)\right)^2 \\ &= p - p^2 \\ &= p(1-p). \end{aligned}$$

Example: Moments from Poisson MGFs

Example (Poisson)

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

$$M_X(t) = \exp\big[\,\lambda(e^t - 1)\,\big].$$

Differentiating this, we obtain

$$M_X'(t) = \lambda e^t M_X(t),$$

and

$$M_X''(t) = \left(\lambda e^t + \lambda^2 e^{2t}\right) M_X(t).$$

Thus,

$$M_X'(0) = \lambda, \qquad M_X''(0) = \lambda + \lambda^2.$$

Hence

$$\mathbb{E}[X] = \lambda, \qquad \text{Var}(X) = \lambda.$$

Example: Moments from the exponential MGF

Example

If $X \sim \text{Exponential}(\lambda)$, then

$$M_X(t) = \frac{\lambda}{\lambda - t}, \qquad t < \lambda.$$

Differentiate:

$$M_X'(t) = \frac{\lambda}{(\lambda - t)^2}, \qquad M_X''(t) = \frac{2\lambda}{(\lambda - t)^3}.$$

Thus,

$$\mathbb{E}[X] = M_X'(0) = \frac{1}{\lambda}, \qquad \mathbb{E}[X^2] = M_X''(0) = \frac{2}{\lambda^2}.$$

Therefore,

$$\text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Pop-up quiz: Moments from an MGF

Suppose a random variable X has MGF

$$M_X(t) = \exp(2t + 3t^2).$$

Question: What are $\mathbb{E}[X]$ and $\mathrm{Var}(X)$?

A) $\mathbb{E}[X] = 2, \quad \mathrm{Var}(X) = 3$
B) $\mathbb{E}[X] = 2, \quad \mathrm{Var}(X) = 6$
C) $\mathbb{E}[X] = 5, \quad \mathrm{Var}(X) = 3$
D) $\mathbb{E}[X] = 5, \quad \mathrm{Var}(X) = 6$

Answer: B. We have

$$M'_X(t) = (2 + 6t)e^{2t+3t^2}, \qquad \text{and thus,} \qquad M'_X(0) = 2.$$

Similarly,

$$M''_X(t) = \left[6 + (2 + 6t)^2\right] e^{2t+3t^2}, \qquad \text{and thus,} \qquad M''_X(0) = 10.$$

Therefore,

$$\mathrm{Var}(X) = M''_X(0) - \left(M'_X(0)\right)^2 = 10 - 4 = 6.$$

MGFs can identify distributions

Inversion property

If $M_X(t)$ is finite for all t in some open interval $(-a, a)$ around 0, then the MGF M_X uniquely determines the distribution (CDF) of X.

If two random variables have the same MGF on an open interval around $t = 0$, then they have the same distribution.

Important remarks:

- The uniqueness statement requires the MGF to exist near 0.
- There exist explicit inverse-transform formulas that allow us to recover the PMF or PDF
- However, we usually identify distributions by "pattern matching" to known MGFs.

Examples: Identifying distributions from MGFs

Example

If

$$M_X(t) = \exp\{\lambda(e^t - 1)\},$$

then $X \sim \text{Poisson}(\lambda)$, because this is the MGF of a Poisson random variable.

Example

If

$$M_X(t) = (1 - p + pe^t)^n,$$

then $X \sim \text{Binomial}(n, p)$, because this is the MGF of a binomial random variable.

To verify this, we can compute the MGF of $Y \sim \text{Binomial}(n, p)$:

$$M_Y(t) = \mathbb{E}[e^{tY}] = \sum_{y=0}^{n} e^{ty}\, p_Y(y) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} p^y (1-p)^{n-y}.$$

Wrap-up

Law of total variance

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

- Average remaining uncertainty plus variability explained by Y

Moment generating functions

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

- If M_X exists near 0, then $M_X^{(k)}(0) = \mathbb{E}[X^k]$
- $\mathrm{Var}(X) = M_X''(0) - \left(M_X'(0)\right)^2$
- MGFs can identify distributions when they exist on an interval around 0

Next lecture: MGFs for sums of independent random variables and random sums

Suggested reading: [BT08, Ch. 4.3 & 4.4]

References


Dimitri Bertsekas and John N Tsitsiklis.
Introduction to probability, volume 1.
Athena Scientific, 2nd edition, 2008.