Chapter 4 and beyond: Learning about distributions from finite data

Part 1, Mon 11 July

Jess Kunke

MATH/STAT 394: Probability I (Summer 2022 A-term)

Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

Announcements + clarifications

- HW 4
- Piecewise function notation

Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

Expectation of a function of a RV

From Chapter 3:

If we know the distribution of a RV X and now we are interested in a RV $Y = g(X)$ for some function g, we know how to compute $\mathrm{E}[Y]$:

Theorem

Let X be a RV that takes values in $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}$ be some function.

$$E[g(X)] = \sum_{k \in \mathcal{X}} g(k)p(k) \qquad \text{if } X \text{ is discrete with pmf } p,$$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx \qquad \text{if } X \text{ is continuous with pdf } f.$$

Now we will cover how to derive the *distribution* of Y from the dist. of X

Invertible functions

Main idea: map values of $Y = g(X)$ back to X

- One concept that might come to mind is the inverse: a map or function $g : A \to B$ is **invertible** if for every $y \in B$ there is a unique $x \in A$ such that $y = g(x)$
- Any monotonic (strictly increasing/decreasing) function g is invertible
- e.g. $g(x) = x^2$ is invertible on $[0, \infty)$

What if g is not invertible?

Maybe multiple values of X map to the same value (y) of Y, so that $g^{-1}(y)$ is a set, not a single number

Images and pre-images of sets

Definition

Let A, B be two sets and $g : A \to B$. The **image** of $F \subseteq A$ under g is defined as

$$g(F) = \{g(x) : x \in F\} \subseteq B.$$

The **pre-image** of $T \subseteq B$ under g is

$$g^{-1}(T) = \{x \in A : g(x) \in T\} \subseteq A.$$

- The notation g^{-1} is the same we use for the inverse of g when it is defined, but here we are not assuming g is invertible
- The pre-image of a set always exists even if the inverse does not exist
- If g **is** invertible, then $g^{-1}(T)$ is the image of T under the inverse map g^{-1}
- These definitions **apply on sets not on variables**
- If there is no element that maps onto T, then $g^{-1}(T) = \emptyset$

Summary: transformations of a discrete RV

Lemma

Let X be a discrete RV, let $g : \mathbb{R} \to \mathbb{R}$, and let $Y = g(X)$. The pmf of Y is

$$p_Y(y) = P(g(X) = y) = P(X \in g^{-1}(\{y\})) = \sum_{\substack{k:g(k)=y \\ k \in \mathcal{X}}} p_X(k).$$

- Why is this result specifically for discrete RVs?

Discrete transformation of a continuous RV

For continuous RVs X let's start by considering the case that the transformation $g(X)$ is discrete.

Example

Suppose that a student's score X is continuous and uniformly distributed on $[0, 100]$: $X \sim \text{Unif}[0, 100]$. A teacher rounds the students' scores to the nearest integer, e.g. if $4.5 \leq X < 5.5$, then the rounded score Y equals 5.
What is the pmf of the rounded scores Y?

Discrete transformation of a continuous RV

More generally we have the following result, which is the same as before:

Lemma

Let X be a continuous RV and $g : \mathbb{R} \to \mathcal{Y}$ be a function that maps $\mathbb{R}$ onto a discrete set $\mathcal{Y}$.
Then the RV Y is discrete and for any $k \in \mathcal{Y}$,

$$P(Y = k) = P(X \in g^{-1}(\{k\})).$$

Summary so far

- If X is discrete and g is any function, then the pmf of $Y = g(X)$ is

$$P(Y = y) = P\big(X \in g^{-1}(\{y\})\big) \quad \text{for } y \in g(\mathcal{X}),$$

where $g^{-1}(\{y\}) = \{k \in \mathcal{X} : g(k) = y\}$ is the pre-image of y under g.

- If X is continuous but g is a discrete function ($g : \mathbb{R} \to \mathcal{Y}$ with $\mathcal{Y}$ discrete), then we have the same result except that we integrate over subsets of $\mathbb{R}$ instead of summing over subsets of a discrete space $\mathcal{X}$:

$$P(Y = y) = P\big(X \in g^{-1}(\{y\})\big),$$

where $g^{-1}(\{y\}) = \{x \in \mathbb{R} : g(x) = y\}$ is the pre-image of y under g.

Continuous transformation of a continuous RV

The cdf method

- For a continuous RV, the pdf has no interpretation as a probability distribution
- It is easier to compute the cdf of $Y = g(X)$, then differentiate to get the pdf
- We will first illustrate the idea, then detail the method if
  1. g is invertible
  2. g is not invertible on $\mathbb{R}$ but invertible on some intervals of $\mathbb{R}$ that form a partition of $\mathbb{R}$

The cdf method: Example

Example

Let X be a continuous RV with pdf f_X. What is the pdf of $Y = aX + b$ for some constants a, b with $a > 0$?

The cdf method: Derivation

g invertible, strictly decreasing

The cdf method: summary

Previous considerations can be summarized by the following lemma:

Lemma

Let X be a continuous RV with pdf f_X.
Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable and strictly monotonic.
Then the pdf of $Y = g(X)$ exists and is given by

$$f_Y(y) = \begin{cases} \left|\frac{1}{g'(g^{-1}(y))}\right| f_X(g^{-1}(y)) & \text{if } y \in g(\mathbb{R}), \\ 0 & \text{otherwise.} \end{cases}$$

Note:

- It is preferable to remember the method rather than the lemma because the method is more flexible (see next slides)

The cdf method: further considerations

1. What if g is not defined on all of $\mathbb{R}$?
   - g only needs to be defined on a subset $B \subseteq \mathbb{R}$ s.t. $P(X \in B) = 1$
   - For $y \notin g(B)$, define $f_Y(y) = 0$

Example

Let $X \sim \text{Unif}([0, 1])$ and $g : x \mapsto -\frac{1}{\lambda} \log(1 - x)$, where $\lambda > 0$. What is the distribution of $Y = g(X)$?

The cdf method: further considerations

2. What if g is not invertible?
   - Partition $\mathbb{R}$ into intervals $[a_i, a_{i+1}]$ such that g is invertible on each interval $[a_i, a_{i+1}]$
   - Apply previous reasoning on these intervals
   - Combine the results to get the cdf, then differentiate to get the pdf

Example

Let X be a continuous RV with pdf f_X. Find the pdf of $Y = X^2$.

The cdf method: further considerations

2. What if g is not invertible?
   - Alternative (ultimately equivalent) approach below

Example

Let X be a continuous RV with pdf f_X. Find the pdf of $Y = X^2$.

Summary: transformations of RVs

- If X is discrete and g is any function, then the pmf of $Y = g(X)$ is

$$P(Y = y) = P\big(X \in g^{-1}(\{y\})\big) \quad \text{for } y \in g(\mathcal{X}),$$

  where $g^{-1}(\{y\}) = \{k \in \mathcal{X} : g(k) = y\}$ is the pre-image of y under g.

- If X is continuous but $Y = g(X)$ is discrete ($g : \mathbb{R} \to \mathcal{Y}$ with $\mathcal{Y}$ discrete), then we have the same result except that we integrate over subsets of $\mathbb{R}$ instead of summing over subsets of a discrete space $\mathcal{X}$:

$$P(Y = y) = P\big(X \in g^{-1}(\{y\})\big),$$

  where $g^{-1}(\{y\}) = \{x \in \mathbb{R} : g(x) = y\}$ is the pre-image of y under g.

- If X and $Y = g(X)$ are continuous, then use the cdf method in some form
  - Identify the possible values of X: only need to account for values of $g(x)$ on this set
  - Identify the possible values of Y: $f_Y(y) = 0$ for all other values in $\mathbb{R}$
  - Compute the cdf of Y
  - If you need the pdf of Y, differentiate

Outline

Motivation

What we've studied so far

- How can we frame a problem in terms of a probability model?
- Given a probability model or probability distribution, how can we compute some useful summaries like the expectation and variance?

But in practice we often don't fully know the distribution

- When you flip a coin, what if you don't want to assume it's fair but instead estimate p?
- What if you want to estimate average income from survey data?
- We can often partially write the probability model but
  - there may be parameters we don't know (e.g. $X \sim \text{Bern}(p)$ but we don't know p)
  - maybe we know the mean and variance but not higher moments

Example: Bernoulli trials

Example

Suppose you have a coin and want to estimate its bias. What would you do?

General setting

- In general, suppose you can frame the problem of interest as a series of independent identically distributed (iid) trials
- A common estimator of the true mean is the **empirical** or **sample mean**

Definition

Let $X_1, \dots X_n \overset{\text{iid}}{\sim} X$ (i.e. n independent RVs, identically distributed as a RV X). The **empirical/sample mean** is defined by

$$\overline{X}_n := \frac{X_1 + \ldots + X_n}{n}.$$

- In the case that some event A can happen or not in each trial and you are interested in the average number of times A happens in n trials, you have $X \sim \text{Binom}(n, p)$ and want to estimate np. What if you know n but not p?
- We will study the empirical mean as a random variable to understand its properties

Properties of the empirical mean of iid trials

- Regardless of whether the X_i are independent, as long as they are identically distributed then we have from linearity of expectation that the mean of the sample mean is the true mean:

$$\mathrm{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_i \mathrm{E}[X_i] = \mathrm{E}[X].$$

We say that the sample mean is **unbiased**.

- If the X_i are iid (why do we need iid here?), then

$$\mathrm{Var}\left(\overline{X}_n\right) = \frac{1}{n^2}\sum_i \mathrm{Var}[X_i] = \frac{1}{n}\mathrm{Var}[X].$$

As the number of trials increases, what happens to the variance of your estimate? Does that make sense?

- Therefore, as $n \to \infty$ it seems that $\overline{X}_n \to \mathrm{E}[X]$. We will develop tools to formalize this (both in this course and in MATH/STAT 395).

Outline

Announcements + clarifications

Distribution of a transformation of a RV (§5.2)

Motivation: Estimation from real data

Concentration inequalities

Estimating tail probabilities: Motivation

- Convergence of the empirical mean will necessarily be stated in terms of probability
- Namely, we would like to show that as $n \to +\infty$, the probability that $\overline{X}_n$ differs from $\mathrm{E}[X]$ tends to 0
- For that we'll need some tools to bound probabilities when we only know e.g. the mean/the variance of a RV
- That's what concentration inequalities are about

First, we'll need the following result:

Theorem (Monotonicity of Expectation)

If two RVs X, Y defined on the same probability space $(\Omega, \mathcal{F}, P)$ have finite means and satisfy that $P(X \leq Y) = 1$, then $E[X] \leq E[Y]$.

Markov's Inequality

What can we say about the probability of X if we know $\mathrm{E}[X]$?

Theorem (Markov's inequality)

If X is a non-negative RV with finite mean, then for any $c > 0$,

$$P(X \geq c) \leq \frac{E[X]}{c}.$$

Proof:

Markov's inequality

Example

A donut vendor sells on average 1000 donuts per day. Could he sell more than 1400 donuts tomorrow with probability greater than 0.8?

Markov's inequality

Example

Let $X \sim \text{Ber}(p)$ for some $p \in (0, 1)$.

1. What is $P(X \geq 0.01)$?
2. What does Markov's inequality give us?

Chebyshev's inequality

What can we say about the probability of X if we know both $E[X]$ and $Var(X)$?

Theorem (Chebyshev's Inequality)

If X is a RV with finite mean μ and finite variance σ^2, then for any $c > 0$,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

Proof:

Note:
The event $\{|X - \mu| \geq c\}$ contains the events $\{X \geq \mu + c\}$ and $\{X \leq \mu - c\}$
So we naturally have a bound on $P(X \geq \mu + c)$, $P(X \leq \mu - c)$

Example

Example

A donut vendor sells on average 1000 donuts per day with a standard deviation of $\sqrt{200}$. Given just this information, provide a bound on

1. the probability that he will sell between 950 and 1050 donuts tomorrow
2. the probability that he will sell at least 1400 donuts tomorrow