Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2023

Recall from last time

- The Joint Distribution Function for $X = (X_1, \ldots, X_n)$ is the function

$$F(x) = \mathbb{P}(X_1 \leq x_1; \ldots; X_n \leq x_n) \quad x = (x_1, \ldots, x_n)$$

Recall from last time

- The Joint Distribution Function for $X = (X_1, \ldots, X_n)$ is the function

$$F(x) = \mathbb{P}(X_1 \leq x_1; \ldots; X_n \leq x_n) \quad x = (x_1, \ldots, x_n)$$

- Random variables $Z = (X, Y)$ are said to be independent if

$$F(z) = F(X \leq x)F(Y \leq y) \quad z = (x, y).$$

Recall from last time

- The Joint Distribution Function for $X = (X_1, \ldots, X_n)$ is the function

$$F(x) = \mathbb{P}(X_1 \leq x_1; \ldots; X_n \leq x_n) \quad x = (x_1, \ldots, x_n)$$

- Random variables $Z = (X, Y)$ are said to be independent if

$$F(z) = F(X \leq x)F(Y \leq y) \quad z = (x, y).$$

- A sequence of random variables $X_1, \ldots, X_n$ is simply a random vector $X = (X_1, \ldots, X_n)$

Recall from last time

- The Joint Distribution Function for $X = (X_1, \dots, X_n)$ is the function

$$F(x) = \mathbb{P}(X_1 \leq x_1; \dots; X_n \leq x_n) \quad x = (x_1, \dots, x_n)$$

- Random variables $Z = (X, Y)$ are said to be independent if

$$F(z) = F(X \leq x)F(Y \leq y) \quad z = (x, y).$$

- A sequence of random variables $X_1, \dots, X_n$ is simply a random vector $X = (X_1, \dots, X_n)$
- The sequence is independent if $F_{X_1,\dots,X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$.

Recall from last time

- The Joint Distribution Function for $X = (X_1, \dots, X_n)$ is the function

$$F(x) = \mathbb{P}(X_1 \leq x_1; \dots; X_n \leq x_n) \quad x = (x_1, \dots, x_n)$$

- Random variables $Z = (X, Y)$ are said to be independent if

$$F(z) = F(X \leq x)F(Y \leq y) \quad z = (x, y).$$

- A sequence of random variables $X_1, \dots, X_n$ is simply a random vector $X = (X_1, \dots, X_n)$
- The sequence is independent if $F_{X_1,\dots,X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$.
- The sequence is identically distributed if $F_{X_i} = F_{X_j}$.

Recall from last time

- The Joint Distribution Function for $X = (X_1, \ldots, X_n)$ is the function

$$F(x) = \mathbb{P}(X_1 \leq x_1; \ldots; X_n \leq x_n) \quad x = (x_1, \ldots, x_n)$$

- Random variables $Z = (X, Y)$ are said to be independent if

$$F(z) = F(X \leq x)F(Y \leq y) \quad z = (x, y).$$

- A sequence of random variables $X_1, \ldots, X_n$ is simply a random vector $X = (X_1, \ldots, X_n)$
- The sequence is independent if $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$.
- The sequence is identically distributed if $F_{X_i} = F_{X_j}$.
- If both then IID (Independent and Identically Distributed)

Learning from data

- What is the average weight of the population in Sweden?

Learning from data

- What is the average weight of the population in Sweden?
- What is a reasonable experiment?

Learning from data

- What is the average weight of the population in Sweden?
- What is a reasonable experiment?
- What is a the random variable? Is it discrete or continuous? Is it bounded?

Learning from data

- What is the average weight of the population in Sweden?
- What is a reasonable experiment?
- What is a the random variable? Is it discrete or continuous? Is it bounded?
- Is our setup of the type IID?

Learning from data

Experiment

Randomly picking a Swedish person and weighing them.

Random variable

X represents the weight of the randomly picked individual. Lets assume that the weight is between 0 and 300. We can state this as $\mathbb{P}(0 \leq X \leq 300) = 1$.

What do we want to learn?

Learning from data

Experiment

Randomly picking a Swedish person and weighing them.

Random variable

X represents the weight of the randomly picked individual. Lets assume that the weight is between 0 and 300. We can state this as $\mathbb{P}(0 \leq X \leq 300) = 1$.

What do we want to learn?

We want to learn $\mathbb{E}[X]$.

Design

How many people should we check the weight of? What do we use to estimate $\mathbb{E}[X]$?

Learning from data

Repeat experiment

Lets now say that we choose to check n people. This is an n-product experiment and we can write the result as $X = (X_1, \ldots, X_n)$ where each X_i is the weight of person i.

Learning from data

Repeat experiment

Lets now say that we choose to check n people. This is an n-product experiment and we can write the result as $X = (X_1, \ldots, X_n)$ where each X_i is the weight of person i.

Estimator

The empirical mean is a good candidate

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{E}[X]?$$

We say that the empirical mean is an **estimator** of $\mathbb{E}[X]$.

Concentration

Concentration of measure

How much is the empirical mean **concentrated** around $\mathbb{E}[X]$?

Concentration

Concentration of measure

How much is the empirical mean **concentrated** around $\mathbb{E}[X]$?

Theorem (**Chebychev's inequality, L^2**)

For **any** *RV X and any* $\epsilon > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathbb{V}(X)}{\epsilon^2}$$

See simulation:

Concentration

Concentration of measure

How much is the empirical mean **concentrated** around $\mathbb{E}[X]$?

Theorem (**Chebychev's inequality, L^2**)

For **any** *RV X and any $\epsilon > 0$,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathbb{V}(X)}{\epsilon^2}$$

See simulation:
Let $\overline{X}_n$ be our empirical mean, and say we choose $\epsilon = 10$, and since $X_i \leq 300$ then $\mathbb{E}[|Z_n|] \leq 300$, so we have

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq 10) \leq \frac{\mathbb{V}[\overline{X}_n]}{100}$$

What does this tell us?

Let $\overline{X}_n$ be our empirical mean, and say we choose $n = 10$ and $\epsilon = 1$, and since $X_i \leq 300$ then $\mathbb{E}[|Z_n|] \leq 300$, so we have

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq 10) \leq \frac{\mathbb{V}[\overline{X}_n]}{100}$$

We have to use our assumptions

1. What can we say about $\mathbb{E}[\overline{X}_n]$?

What does this tell us?

Let $\overline{X}_n$ be our empirical mean, and say we choose $n = 10$ and $\epsilon = 1$, and since $X_i \leq 300$ then $\mathbb{E}[|Z_n|] \leq 300$, so we have

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq 10) \leq \frac{\mathbb{V}[\overline{X}_n]}{100}$$

We have to use our assumptions

1. What can we say about $\mathbb{E}[\overline{X}_n]$?
2. What can we say about $\mathbb{V}[\overline{X}_n]$?

What does this tell us?

Let $\overline{X}_n$ be our empirical mean, and say we choose $n = 10$ and $\epsilon = 1$, and since $X_i \leq 300$ then $\mathbb{E}[|Z_n|] \leq 300$, so we have

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq 10) \leq \frac{\mathbb{V}[\overline{X}_n]}{100}$$

We have to use our assumptions

1. What can we say about $\mathbb{E}[\overline{X}_n]$?
2. What can we say about $\mathbb{V}[\overline{X}_n]$?

That is

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq 10) \leq \frac{\mathbb{V}[X]}{100n}$$

What does this mean?

Building confidence intervals

That is

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq 10) \leq \frac{\mathbb{V}[X]}{100n}$$

What does this mean?

- We can use the statement to build confidence intervals.

Building confidence intervals

That is

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq 10) \leq \frac{\mathbb{V}[X]}{100n}$$

What does this mean?

- We can use the statement to build confidence intervals.
- If we want the probability of our measurement landing within 10 from the true expectation to be larger than 95% we need to find n such that

$$\frac{\mathbb{V}[X]}{100n} \leq 1 - 0.95$$

Building confidence intervals

That is

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq 10) \leq \frac{\mathbb{V}[X]}{100n}$$

What does this mean?

- We can use the statement to build confidence intervals.
- If we want the probability of our measurement landing within 10 from the true expectation to be larger than 95% we need to find n such that

$$\frac{\mathbb{V}[X]}{100n} \leq 1 - 0.95$$

- $0 \leq X \leq 300$ implies $\mathbb{V}[X] \leq 300^2/4 = 22500$, thus we need $22500/5 = 4500$ samples.

Confidence interval

If we look at the statement from before

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

Confidence interval

If we look at the statement from before

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

If we use the complementary event we get

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| < \epsilon) \geq 1 - \delta$$

Confidence interval

If we look at the statement from before

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

If we use the complementary event we get

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| < \epsilon) \geq 1 - \delta$$

Now by rearranging we get

$$\mathbb{P}(\overline{X}_n - \epsilon < \mathbb{E}[X] < \overline{X}_n + \epsilon) \geq 1 - \delta$$

Confidence interval

If we look at the statement from before

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

If we use the complementary event we get

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| < \epsilon) \geq 1 - \delta$$

Now by rearranging we get

$$\mathbb{P}(\overline{X}_n - \epsilon < \mathbb{E}[X] < \overline{X}_n + \epsilon) \geq 1 - \delta$$

Confidence interval

For this example the confidence interval is the interval

$$(\overline{X}_n - \epsilon, \overline{X}_n + \epsilon).$$

Confidence interval

Confidence interval

If we let

$$I = (\overline{X}_n - \epsilon, \overline{X}_n + \epsilon) \quad \text{then} \quad \mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta.$$

Which of the following is true?

1. Lets say we used data and got an interval of $(0.1, 0.3)$, then the probability that the confidence interval contains the expectation is greater than $1 - \delta$.

Confidence interval

Confidence interval

If we let

$$I = (\overline{X}_n - \epsilon, \overline{X}_n + \epsilon) \quad \text{then} \quad \mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta.$$

Which of the following is true?

1. Before we have computed the interval with data, the probability that the random interval contains $\mathbb{E}[X]$ is greater than or equal to $1 - \delta$.

Confidence interval

Confidence interval

If we let

$$I = (\overline{X}_n - \epsilon, \overline{X}_n + \epsilon) \quad \text{then} \quad \mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta.$$

Which of the following is true?

1. If I repeat the experiment of collecting data and each time computing the confidence interval then I should see roughly $1 - \delta$ or more of them containing $\mathbb{E}[X]$.

Confidence interval

Confidence interval

If we let

$$I = (\overline{X}_n - \epsilon, \overline{X}_n + \epsilon) \quad \text{then} \quad \mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta.$$

Which of the following is true?

1. Before I repeat the experiment of collecting data and each time computing the confidence interval, I expect to see roughly $1 - \delta$ or more of them containing $\mathbb{E}[X]$.

Confidence interval

Confidence interval

If we let

$$I = (\overline{X}_n - \epsilon, \overline{X}_n + \epsilon) \quad \text{then} \quad \mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta.$$

Which of the following is true?

1. If I in the future will compute confidence intervals with $1 - \delta$ for the rest of my professional life, then I will produce intervals covering the true expectation roughly $1 - \delta$ or more of the time.

Can we do better?

Theorem (Hoeffdings inequality)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability triple and let* $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ *be* $\mathbb{R}$*-valued RVs such that* $\mathbb{P}(X_i \in [a, b]) = 1$*, then for any* $\epsilon > 0$ *we get for* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$*,*

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Can we do better?

Theorem (Hoeffdings inequality)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\mathrm{IID}}{\sim} F$ be $\mathbb{R}$-valued RVs such that $\mathbb{P}(X_i \in [a, b]) = 1$, then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

See simulation:

Can we do better?

Theorem (Hoeffdings inequality)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X_1, \ldots, X_n \overset{\text{IID}}{\sim} F$ be $\mathbb{R}$-valued RVs such that $\mathbb{P}(X_i \in [a, b]) = 1$, then for any $\epsilon > 0$ we get for $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

See simulation:
Again choose $\epsilon = 10$ and find n such that

$$2e^{-\frac{2n}{900}} = 0.05$$

The solution is given by

$$1700 \approx 450 * \ln(1/0.025) = n$$

Can we do better still?

Assumptions

If we make no further assumptions, we cannot do better!!

Can we do better still?

Assumptions

If we make no further assumptions, we cannot do better!!

Small variance

It is not unreasonable to think that the variance is not as big as $300^2/4$ as that would correspond to half the population having weight 0 and the other half having weight 300.

Can we do better still?

Assumptions

If we make no further assumptions, we cannot do better!!

Small variance

It is not unreasonable to think that the variance is not as big as $300^2/4$ as that would correspond to half the population having weight 0 and the other half having weight 300.

Data

I could not find any weight data, but I could find some data on BMI instead. Here the variance is roughly 34.

Assumptions

- For simple output like Bernoulli we get good bounds.
- For random variables with a large span, it is often better to use some guided assumptions about either ”spread” or how heavy the tails are.

Tail assumptions

For random variables with large range but has a small spread, we can use the following instead

Theorem (Bennett's inequality)

Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite variance such that $\mathbb{P}(X_i \leq b) = 1$ with mean zero. Let and $\sigma^2 = \mathbb{V}[X_i]$. Then for any $\epsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2 \exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon}{\sigma^2}\right)\right)$$

where $h(u) = (1+u)\log(1+u) - u$ for $u > 0$.

Going back to our example of measuring weight, if we assume that $\sigma = 20$ we get that n should be roughly 50 for $\epsilon = 10$.