Analysis Methods for Cross-Sectional Data: Probability and Statistics

Tatiana Zolotareva

May 8, 2019

Probability

1. Alice has 2 kids, her first child is a girl. Find the probability that the second child is also a girl.

2. Alice has 2 kids, one of them is a girl. Find the probability that both of them are girls.

3. Monty Hall Problem: Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Probability

Answers:

1. A = { First child is a girl }
   B = { Second child is a girl }
   $A \cap B$ = {Both kids are girls}

   $$P(B|A) = P(A \cap B)/P(A) = P(A)P(B)/P(A) = P(B) = \frac{1}{2}$$

2. C = { One of the kids is a girl }

   $$P(A \cap B|C) = P(A \cap P \cap C)/P(C) = P(A \cap B)/P(C) = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

Probability

Answers

3. A = {Car is behind door 1},
   B = {Car is behind door 2},
   C = {Car is behind door 3},
   D = {Host shows door 3 after you pick door 1}

   $$P(A) = \frac{1}{3}, \quad P(B) = \frac{1}{3}, \quad P(C) = \frac{1}{3}, \quad P(D) = \frac{1}{2}$$

   $$P(D|C) = 0, \quad P(D|A) = \frac{1}{2}, \quad P(D|B) = 1$$

   $$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{1}{3}$$

   $$P(B|D) = \frac{P(D|B)P(B)}{P(D)} = \frac{2}{3}$$

Probability

Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Independent events:
$P(A|B) = P(A) \iff P(A \cap B) = P(A)P(B)$

Bayes Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Likelihood:
$P(target|input) \propto P(input|target)P(target)$
$P(target|input)$ - Posterior probability
$P(input|target)$ - Likelihood
$P(target)$ - Prior probability

Probability

Definition

Probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of:

- Sample space Ω of all possible outcomes of the experiment
- σ-algebra of events $\mathcal{F}$
- A probability measure P that assigns probability to events

Probability

Definition

σ-**algebra** is a collection of sets of outcomes in Ω such that:

1. $S \in \mathcal{F} \implies S^c \in \mathcal{F}$
2. $\Omega \in \mathcal{F}$ (an event of all possible outcomes)
3. $S_1, S_2, \ldots \in \mathcal{F} \implies \bigcup S_i \in \mathcal{F}$

Experiment: Flip a coin once.
$\Omega =$

Probability

Definition

σ-**algebra** is a collection of sets of outcomes in Ω such that:

1. $S \in \mathcal{F} \implies S^c \in \mathcal{F}$
2. $\Omega \in \mathcal{F}$ (an event of all possible outcomes)
3. $S_1, S_2, \ldots \in \mathcal{F} \implies \bigcup S_i \in \mathcal{F}$

Experiment: Flip a coin once.
$\Omega = \{\text{heads, tails}\}$
$\mathcal{F} =$

Probability

Definition

σ-**algebra** is a collection of sets of outcomes in Ω such that:

1. $S \in \mathcal{F} \implies S^c \in \mathcal{F}$
2. $\Omega \in \mathcal{F}$ (an event of all possible outcomes)
3. $S_1, S_2, \ldots \in \mathcal{F} \implies \bigcup S_i \in \mathcal{F}$

Experiment: Flip a coin once.
$\Omega = \{\text{heads, tails}\}$
$\mathcal{F} = \{\text{heads, tails, heads or tails, neither heads nor tails}\}$

Probability

Definition

Probability measure is a function on $\mathcal{F}$ such that

1. $P(S) >= 0$ for any $S \in \mathcal{F}$
2. $P(\Omega) = 1$
3. $P(\bigcup_{i=1}^{n} S_i) = \sum_{i=1}^{n} S_i, \quad \lim_{n\to\infty} P(\bigcup_{i=1}^{n} S_i) = \sum_{i=1}^{\infty} S_i$

Random variables

Definition

A random variable X is a measurable function from the space of possible outcomes Ω to $\mathbb{R}$.

Example: Throwing two dice, X is the obtained score. There are multiple outcomes that yield a score.

Events $\{\text{outcomes } w \in \Omega : X(w) \in I\}$ for all intervals $I \subset \mathbb{R}$ form a σ-algebra.

Discrete Random Variables

A discrete random variable is defined over a discrete space of outcomes $X : \Omega \rightarrow \mathbb{D}_X$. The **probability mass function** is then defined by:

$$p_X(x) = P(X = x) = P(\{w \in \Omega \mid X(w) = x\})$$

$$\sum_{x \in \mathbb{D}_X} p_X(x) = 1$$

Example: Let X be the random variable representing the score in the experiment of throwing two dice once. Then,

$$\begin{aligned} p_X(4) &= P(X = 4) \\ &= P(\{\text{dice } 1 = 2, \text{ dice } 2 = 2\} \text{ or} \\ &\qquad \{\text{dice } 1 = 3, \text{ dice } 2 = 1\} \\ &\qquad \{\text{dice } 1 = 1, \text{ dice } 2 = 3\}) = 3\frac{1}{36} = \frac{1}{9} \end{aligned} \tag{1}$$

Discrete Random Variables

If X and Y are random variables on the same probability space, then the **joint probability mass function** is defined as:

$$p_{X,Y}(x,y) = P(\{w \in \Omega \mid X(w) = x \text{ and } Y(w) = y\})$$

and verifies the properties:

$$\sum_{x \in \mathbb{D}} p_{X,Y}(x,y) = p_Y(y), \quad \sum_{y \in \mathbb{D}} p_{X,Y}(x,y) = p_X(x)$$

Independent random variables:

$$p_{X,Y}(x,y) = p_X(x) p_Y(y)$$

Conditional probability mass function:

$$p_{X|Y}(x|y) = P(X = x | Y = y) = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

Discrete Random Variables

Definition

Expected value of a function $f : \mathbb{R} \to \mathbb{R}$ is defined by:

$$\mathbb{E}[f(X)] = \sum_{x \in \mathbb{D}_X} f(x) p_X(x)$$

$$\mathbb{E}[X] = \sum_{x \in \mathbb{D}_X} x \; p_X(x)$$

Exercise 1: Show that expectation is linear:

$$\mathbb{E}\left[\alpha X + \beta Y\right] = \alpha[X] + \beta[Y]$$

Exercise 2: Show that for independent random variables X, Y:

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right] \mathbb{E}\left[Y\right]$$

Discrete Random Variables

Solution 1: Consider a tuple of random variables (X, Y) and a function $f : x, y \mapsto \alpha x + \beta y$. By definition:

$$\begin{aligned}
\mathbb{E}\left[\alpha X + \beta Y\right] &= \sum_{x \in \mathbb{D}_X} \sum_{y \in \mathbb{D}_Y} (\alpha x + \beta y) p_{X,Y}(x, y) \\
&= \alpha \sum_{x \in \mathbb{D}_X} x \sum_{y \in \mathbb{D}_Y} p_{X,Y}(x, y) + \beta \sum_{y \in \mathbb{D}_Y} y \sum_{x \in \mathbb{D}_X} p_{X,Y}(x, y) \\
&= \alpha \sum_{x \in \mathbb{D}_X} x p_X(x) + \beta \sum_{y \in \mathbb{D}_Y} y p_Y(y) \\
&= \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]
\end{aligned}$$

Discrete Random Variables

Solution 2: Since X and Y are independent, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. Then,

$$\mathbb{E}[X, Y] = \sum_{x \in \mathbb{D}_X} \sum_{y \in \mathbb{D}_Y} xyp_X(x)p_Y(y)$$
$$= \sum_{x \in \mathbb{D}_X} xp_X(x) \sum_{y \in \mathbb{D}_y} yp_X(y) = \mathbb{E}[X]\mathbb{E}[Y]$$

Discrete Random Variables

Definition

Variance of a random variable X is defined as:

$$var(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2$$

Covariance of two random variables X and Y is defined as:

$$cov(X, Y) = \mathbb{E}\Big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\Big] = \mathbb{E}\left[XY\right] - \mathbb{E}[X]\mathbb{E}[Y]$$

Exercise 3: Show that for two independent random variables X and Y

$$var(X + Y) = var(X) + var(Y)$$

Discrete Random Variables

Solution 3: Since X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, then

$$\begin{aligned}
var(X+Y) &= \mathbb{E}\left[(X+Y)^2\right] - (\mathbb{E}\left[X+Y\right])^2 \\
&= \mathbb{E}[(X^2 + 2XY + Y^2)] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y^2] \\
&- (E[X])^2 - (E[X])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] \\
&= var(X) + var(Y)
\end{aligned}$$

Important probability distributions

Example 1: The probability to win 10 dollars in the lottery is 0.001. What is my expected gain if I play only once?

Important probability distributions

Example 1: The probability to win 10 dollars in the lottery is 0.001. What is my expected gain if I play only once?

$$\begin{aligned}\mathbb{E}[gain(X)] &= 10 * P(win) + 0 * P(loose)\\ &= 10 \times 0.001 = 0.01 \text{ dollars}\end{aligned}$$

Bernoulli distribution with parameter $p \in [0, 1]$:

$P(X = 1) = p, \quad P(X = 0) = 1-p, \quad \mathbb{E}[X] =$

Important probability distributions

Example 1: The probability to win 10 dollars in the lottery is 0.001. What is my expected gain if I play only once?

$$\begin{aligned}\mathbb{E}[gain(X)] &= 10 * P(win) + 0 * P(loose)\\ &= 10 \times 0.001 = 0.01 \text{ dollars}\end{aligned}$$

Bernoulli distribution with parameter $p \in [0, 1]$:

$P(X = 1) = p, \quad P(X = 0) = 1-p, \quad \mathbb{E}[X] = p, \quad var(X) =$

Important probability distributions

Example 1: The probability to win 10 dollars in the lottery is 0.001. What is my expected gain if I play only once?

$$\begin{aligned}\mathbb{E}[gain(X)] &= 10 * P(win) + 0 * P(loose) \\ &= 10 \times 0.001 = 0.01 \text{ dollars}\end{aligned}$$

Bernoulli distribution with parameter $p \in [0, 1]$:

$$P(X = 1) = p, \quad P(X = 0) = 1-p, \quad \mathbb{E}[X] = p, \quad var(X) = p-p^2$$

Important probability distributions

Example 2: If play the lottery more than once my chances to win are better. How probabal is it that I will need to buy at least 100 tickets before I win? What is the expected number of tickets I need to buy before I win?

Important probability distributions

Example 2: If play the lottery more than once my chances to win are better. How probabal is it that I will need to buy at least 100 tickets before I win? What is the expected number of tickets I need to buy before I win?

$p(\text{loose 100 times before win}) = (0.999)^{99} * 0.001 \sim 0.0009$

Geometric distribution with parameter p:

$$P(x = k) = (1 - p)^{k-1} p, \quad k = 1, 2, \ldots$$

$$\mathbb{E}[X] = \frac{1}{p}, \quad var(X) = \frac{1-p}{p^2}$$

Important probability distributions

In order to calculate the expectation and the variance, we calculate infinite sums:

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} kp(1-p)^{k-1} = -p\frac{d}{dp}\sum_{k=0}^{\infty}(1-p)^k$$
$$= -p\frac{d}{dp}\left(\frac{1}{p}\right) = \frac{p}{p^2} = \frac{1}{p}$$

$$var(X) = \sum_{k=0}^{\infty} k^2 p(1-p)^{k-1} - \frac{1}{p} = \frac{1-p}{p^2}$$

Trick: represent the series as derivatives of well-known series

Important probability distributions

Example 3: I would like to win at least 20 dollars. How many tickets do I need to buy?

Important probability distributions

Example 3: I would like to win at least 20 dollars. How many tickets do I need to buy?

$$\mathbb{E}[gain(\text{k tickets})] = k * \mathbb{E}[gain(\text{1 ticket})] = k * 0.01 = 20 \implies k = 2000$$

What is the probability that I win at least twice with 2000 tickets?
What is the probability that I win exactly twice?

Important probability distributions

Example 3: I would like to win at least 20 dollars. How many tickets do I need to buy?

$$\mathbb{E}[gain(\text{k tickets})] = k*\mathbb{E}[gain(1 \text{ ticket})] = k*0.01 = 20 \implies k = 2000$$

What is the probability that I win at least twice with 2000 tickets?
What is the probability that I win exactly twice?

$$\begin{aligned} P(\text{win at least twice}) &= 1 - p(\text{never win}) - p(\text{win once}) \\ &= 1 - (0.999)^{2000} - 1000 \times (0.999)^{1999} \times 0.001 = 0.5943 \end{aligned}$$

$$p(\text{win exactly twice}) = \binom{2000}{2} \times 0.001^2 \times 0.999^{1998} = 0.2708$$

Binomial distribution: with parameter p:

$$P(X = k|n) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n$$

$$\mathbb{E}[X] = np, \; var(X) = np(1-p)$$

Important probability distributions

Example 4: In total 1000 lottery tickets were printed: 10 of them are win tickets. Today 200 tickets were sold. What is the probability that exactly 2 of them would be win tickets?

Important probability distributions

Example 4: In total 1000 lottery tickets were printed: 10 of them are win tickets. Today 200 tickets were sold. What is the probability that exactly 2 of them would be win tickets?

$$P(X=2) = \frac{\binom{10}{2}\binom{990}{198}}{\binom{1000}{200}} = 0.3$$

Important probability distributions

Example 4: In total 1000 lottery tickets were printed: 10 of them are win tickets. Today 200 tickets were sold. What is the probability that exactly 2 of them would be win tickets?

$$P(X = 2) = \frac{\binom{10}{2}\binom{990}{198}}{\binom{1000}{200}} = 0.3$$

Hypergeometric distribution with parameters $N = 1000$-population size, $n = 200$-number of draws without replacement, $K = 10$-number of successes in the population, $k = 2$-number of observed successes:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$\mathbb{E}[X] = np, \quad var(X) = \frac{np(1-p)(N-n)}{N-1}, \quad p = \frac{K}{N}$$

Exercise: Confirm the result using `scipy.stats.hypergeom`.

Important probability distributions

Example 5: What is the probability that k people will win the lottery today if in average λ people win the lottery in one day?

Important probability distributions

Example 5: What is the probability that k people will win the lottery today if in average λ people win the lottery in one day?

Assume day $= n$ time intervals
only one lottery ticket can be won in one time interval
$p(\text{winner in one time interval}) = \frac{\lambda}{n}$, $n \to \infty$

$$P(k \text{ winners during the day}) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$P(k \text{ winners in t days}) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

Poisson distribution with parameter λ:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

$$\mathbb{E}[X] = \lambda, \quad var(X) = \lambda^2 + \lambda$$

Summary Discrete Random Variables

1. One trial with binary outcome: success, failure $\Longrightarrow$ **Bernoulli distribution**
2. Number of trials till success $\Longrightarrow$ **Geometric distribution**
3. Number of successes in n trials $\Longrightarrow$ **Binomial distribution**
4. Given the number of successes in the population, find number of successes in a sample (drawn without replacement) $\Longrightarrow$ **Hypergeometric distribution**
5. Number of successes in a time period, given the average number of successes per time period $\Longrightarrow$ **Poisson distribution**

Continuous random variables

Probability density function: Probability that value of the random variable X belongs to the interval Δx ($|\Delta x| \to 0$) can be approximated by $p(x)\Delta x$.

$$P(X \in (a,b]) = \int_a^b p(x)dx, \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

Cumulative distribution function:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(x)dx$$

Expected value of a function $f : \mathbb{R} \to \mathbb{R}$ of random variable X:

$$\mathbb{E}\left[f(X)\right] = \int_{-\infty}^{\infty} f(x)p(x)dx, \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx$$

Continuous random variables

Joint CDF of random variables X and Y:

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p_{X,Y}(x, y) dx dy$$

$$\int_{-\infty}^{\infty} p_{X,Y}(x, y) dy = p_X(x), \quad \int_{-\infty}^{\infty} p_{X,Y}(x, y) dx = p_Y(y)$$

Conditional Distributions:

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

If X and Y are independent random variables,

$$p_{Y|X}(y|x) = p_Y(y) \iff p_{X,Y}(x, y) = p_X(x) p_Y(y)$$

Important Continuous Distributions

Example 1: The bus leaves the bus-stop every 15 minutes. What is the probability that you will wait less than 5 minutes for the next bus? How much time are you going to wait in average?

Important Continuous Distributions

Example 1: The bus leaves the bus-stop every 15 minutes. What is the probability that you will wait less than 5 minutes for the next bus? How much time are you going to wait in average?

The probability to wait less than x minutes increases at a constant speed when x increases.

$$P(X \leq x) = cx, \quad \int_0^{15} c dx = 1 \implies c = \frac{1}{15}, \quad P(X \leq 5) = \frac{1}{3}$$

Uniform distribution over the interval (a, b):

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

Important Continuous Distributions

Example 1: The bus leaves the bus-stop every 15 minutes. What is the probability that you will wait less than 5 minutes for the next bus? How much time are you going to wait in average?

The probability to wait less than x minutes increases at a constant speed when x increases.

$$P(X \leq x) = cx, \quad \int_0^{15} c dx = 1 \implies c = \frac{1}{15}, \quad P(X \leq 5) = \frac{1}{3}$$

Uniform distribution over the interval (a, b):

$$p(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{2}(a+b), \quad var(X) = \frac{1}{12}(b-a)^2$$

Important Continuous Distributions

Example 2: You observe that the number of hits on your web-site follows a Poisson distribution at the rate 2 per day. What is the probability that you will have to wait less than 5 days until the next hit.

Important Continuous Distributions

Example 2: You observe that the number of hits on your web-site follows a Poisson distribution at the rate 2 per day. What is the probability that you will have to wait less than 5 days until the next hit.

$$P(X > t) = P(0 \text{ hits in t days}) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

$$P(X \leq 5) = 1 - P(x > 5) = 1 - e^{-5\lambda} = 1 - e^{-10}$$

Exponential distribution with parameter λ:

$$P(X \leq x) = 1 - e^{-\lambda x}, \quad p(x) = \lambda e^{-\lambda x}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}, \quad var(X) = \frac{2}{-\lambda^2}$$

Important Continuous Distributions

Normal (Gaussian) distribution with mean μ and standard deviation σ:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt$$

Important Continuous Distributions

Normal (Gaussian) distribution with mean μ and standard deviation σ:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt$$

Theorem

Central Limit Theorem: *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $var(X_i) = \sigma^2 < \infty$, then*

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1) \text{ in distribution}$$

Summary Continuous Random Variables

1. If Probability of a random variable to belong to an interval grows linearly when the interval grows $=>$ **uniform distribution**
2. Given the average number of successes per time unit, time until success $=>$ **Exponential distribution**
3. Random variable representing an average value in a sample $=>$ **Normal distribution**

Sample mean and variance

What do we do if we have observations (data) but do not know the distribution followed by the data?

Law of large numbers: Let $X_1, X_2, \dots,$ be a sequence of i.i.d. random variables with the expected value μ and variance σ^2. Then, the sample mean and the sample variance defined as

$$\mu_n = \frac{\sum_{k=1}^{n} x_k}{n}, \quad \sigma_n^2 = \frac{\sum_{k=1}^{n} (x_k - \mu_n)^2}{n}$$

converge in probability to the mean and the variance:

$$\mu_n \xrightarrow[n\to\infty]{} \mu, \quad \sigma_n^2 \xrightarrow[n\to\infty]{} \sigma^2$$

Sample mean and variance

$$\mathbb{E}[\sigma_n^2] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\left(x_k - \frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right]$$

$$= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\left((x_k-\mu) - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)\right)^2\right]$$

$$= \frac{1}{n}\sum_{k=1}^{n}\left(\mathbb{E}\left[(x_k-\mu)^2\right] - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[(x_k-\mu)(x_i-\mu)\right]\right.$$

$$\left. + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(x_i-\mu)^2\right]\right) = \frac{n-1}{n}\sigma^2$$

Unbiased sample variance:

$$\tilde{\sigma}_n^2 = \sum_{k=1}^{n}\frac{(x_k-\mu_n)^2}{n-1}, \quad \mathbb{E}\left[\tilde{\sigma}_n^2\right] = \sigma^2$$

Fitting a probability distribution to data

1. Choose the familty of probability distributions
2. Find the parameters of the distribution that maximize the likelihood of obtaining the data

$$P(x_1, \ldots, x_n | parameters) \rightarrow max$$

Maximum Likelihood Estimates

Example: $x_1, \ldots, x_n$ - realizations of a **Bernoulli** random variable. Estimate the parameter p - probability of success.

$$P(x_1, \ldots, x_n | p) = \prod_{k=1}^{n} P(x_k | p) = \prod_{k=1}^{n} p^{x_k} (1-p)^{1-x_k}$$

$$f(p) = \log P(x_1, \ldots, x_n | p) = \sum_{k=0}^{n} \left(x_k \log p + (1 - x_k) \log(1-p) \right)$$

$$\frac{df}{dp} = \sum_{k=0}^{n} \left(\frac{x_k}{p} - \frac{1-x_k}{1-p} \right) = 0 \quad \Longleftrightarrow \quad p_{ML} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\frac{d^2 f}{dp^2} < 0$$

Conclusion: f is a convex function and attains its maximum when the probaility of success is approximated by the ratio of the successes in the sample data (sample mean).

Maximum Likelihood Estimates

Poisson, Exponential distribution:

$$\lambda_{ML} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

Normal distribution:

$$\mu_{ML} = \frac{1}{n} \sum_{k=1}^{n} x_k, \quad \sigma_{ML}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu_{ML})^2$$

Statistical Significance

Usecase: You have an online shop and you pay facebook to show your add. You suspect that more people navigate to your web-site in the second part of the day. You make observations during 20 days and in 14 cases your conjecture confirmed.

- Would you ask facebook to make more impressions of your add in the second part of day?

Statistical Significance

Usecase: You have an online shop and you pay facebook to show your add. You suspect that more people navigate to your web-site in the second part of the day. You make observations during 20 days and in 14 cases your conjecture confirmed.

- Would you ask facebook to make more impressions of your add in the second part of day?
- What if your conjecture was true in 1400 out of 2000 observations?

Statistical Significance

Techniques for evaluating a pre-defined conjecture is called **hypothesis testing**.

H_0 - **null hypothesis**: hypothesis that our conjecture is false
H_1 - **alternative hypothesis**: hypothesis under which our conjecture is true

Type I error: conjecture is false, but H_0 is rejected
Type II error: conjecture is true, but H_0 is not rejected

What type of error the hypothesis testing aims to avoid?

Statistical Significance

Definition

p-value is the probability (likelihood) to observe results at least as extreme as those measured under the assumption that the null hypothesis H_0 is true.

Wrong interpretation: probability that the null hypothesis is true.

The null-hypothesis H_0 is rejected if the p-value is less than a given **significance level** and the measured data is called statistically significant.

Significance level has to be decided on. (Popular choices 5%, 1%)

Statistical Significance: Binomial Test

$H_0 =$ number of conversions does not depend on the time of the day

$$X = \begin{cases} 1, & \text{if there more conversions in the second part of the day} \\ 0, & \text{otherwise} \end{cases}$$

Under H_0: $p(X = 1) = \frac{1}{2}$

$t = \sum_{i=1}^{n} X_i$ - number of days with more conversions in the second part of the day

$p(t >= k) =$

Statistical Significance: Binomial Test

$H_0 =$ number of conversions does not depend on the time of the day

$$X = \begin{cases} 1, & \text{if there more conversions in the second part of the day} \\ 0, & \text{otherwise} \end{cases}$$

Under H_0: $p(X = 1) = \frac{1}{2}$

$t = \sum_{i=1}^{n} X_i$ - number of days with more conversions in the second part of the day

$p(t >= k) = \frac{1}{2^n} \sum_{i=k}^{n} \binom{n}{k}$

Exercise: calculate in python for n=20, k=14 (n=2000, k=1400) using function `scipy.special.binom`. Confirm the results with the function `scipy.stats.binom_test`.

Statistical Significance

Now let's say that the online-shop considers to rearrange the landing page of its web-site. The metrics monitored by the shop are:

- Average time spent on the landing page per session
- Conversion rate = average proportion of sessions that end up with a transaction.

  How are you going to evaluate if the changes in the landing page increase the income?

Statistical Significance

You split the traffic of the web-site randomly between two site versions in proportion 60%-40%:

Version	n sessions	avg(time)	stdtime	number of conversions
A	6000	60s	40s	90
B	4000	62s	45s	80

Statistical Significance

You split the traffic of the web-site randomly between two site versions in proportion 60%-40%:

Version	**n sessions**	**avg(time)**	**stdtime**	**number of conversions**
A	6000	60s	40s	90
B	4000	62s	45s	80

$$CR(A) = \frac{90}{6000} = 0.015, \quad CR(B) = \frac{80}{4000} = 0.02$$

Which web-site version is better in terms of time spent on the page and the conversion rate?

Statistical Significance

A/B versions == treatement/control groups

Use-cases for A/B testing:

- Product or service development
- Medicine (to test effects of a treatment)
- In economics (to undersand the behavior of economical actors)

Important Aspects:

- Randomization strategy: There should be no hidden factors that bias the the selection. Example: selling two versions of a product in two shops with different geo-locations. (A solution: increase the number of shops)
- Sample size should be sufficient to represent the population

Statistical Significance

$\mathbf{H_0}$: Time spent on the landing page does not depend on the page version.

$\mathbf{H_1}$: Time spent on the landing page depends on the page version.

- What distribution does the mean time spent on the web-site follow?
- What statistic should we consider?

Statistical Significance: Normal test

$$\hat{t}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} t_i \sim \mathcal{N}(\mu_A, \frac{\sigma_A}{\sqrt{n_A}}), \quad \hat{t}_B = \frac{1}{n_B}\sum_{i=1}^{n_B} t_i \sim \mathcal{N}(\mu_B, \frac{\sigma}{\sqrt{n_B}})$$

Under H_0: $\quad \hat{t}_A - \hat{t}_B \sim \mathcal{N}(0, \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}})$

$$Z = \frac{\hat{t}_A - \hat{t}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim \mathcal{N}(0, 1)$$

Exercise: Find the corresponding p-value using `scipy.stats.norm`. Plot the function `scipy.stats.norm.cdf`.

Statistical Significance: T-test

When the number of observations is very large, normal distribution is not a good approximation for the Z statistic. In this case, we use the Student's T-distribution:

$$T = \frac{\hat{t}_A - \hat{t}_B}{\sigma_{pooled}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim \mathcal{T}(0, n_A + n_B - 1)$$

$$\sigma_{pooled} = \frac{(n_A - 1)\sigma_A^2 + (n_B - 1)\sigma_B^2}{n_A + n_B - 2}$$

follows the Student's t-distribution with $n_A + n_B - 1$ degrees of freedom.

Exercise: Find the corresponding p-value using `scipy.stats.t`.

Statistical Significance: $\chi 2$ Test

Let's say you distribute n objects in r boxes with the probability to arrive in the box B_j equal to p_j. And let ν_j be a random variable that describes the number of objects in the box B_j. Than,

$$\mathbb{E}(\nu_j) = np_j, \quad var(\nu_j) = np_j(1 - p_j), \quad \frac{\nu_j - np_j}{\sqrt{npj(1 - p_j)}} \to \mathcal{N}(0, 1)$$

Pearson's Theorem:

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \to \chi^2_{r-1}$$

convergence in distribution to χ^2_{r-1} with $r - 1$ degrees of freedom.

$$\sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

Statistical Significance: $\chi 2$ Test

If now you randomly distribute n objects of k colors and r boxes, let ν_{ij} be the number of objects of color i in the box B_j. Let probability to arrive in the box B_j is equal to p_j and the probability to pick an object of color i is equal to q_i. Then,

Pearson's Theorem:

$$\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(\nu_{ij} - np_j q_i)^2}{np_j q_i} \to \chi^2_{(r-1)\times(k-1)}$$

convergence in distribution to $\chi^2_{(r-1)\times(k-1)}$ with $(r-1)\times(k-1)$ degrees of freedom.

Statistical Significance: $\chi 2$ Test

$\mathbf{H_0}$: The number of conversions does not depend on the page version.

"Boxes" - versions A and B of the web-site, with probabilities:

Statistical Significance: $\chi 2$ Test

$\mathbf{H_0}$: The number of conversions does not depend on the page version.

"Boxes" - versions A and B of the web-site, with probabilities:

$$p_A = 0.6, \quad p_B = 0.4$$

"Colors" - conversion, no-conversion, with probabilities (under H_0):

Statistical Significance: $\chi 2$ Test

$\mathbf{H_0}$: The number of conversions does not depend on the page version.

"Boxes" - versions A and B of the web-site, with probabilities:

$$p_A = 0.6, \quad p_B = 0.4$$

"Colors" - conversion, no-conversion, with probabilities (under H_0):

$$q_c = \frac{80 + 90}{6000 + 4000} = 0.017, \quad q_{nc} = 1 - q_c = 0.983, \quad n = 1000$$

$$\nu_{11} = 90, \quad \nu_{12} = 6000 - 90 = 5910$$

$$\nu_{21} = 80, \quad \nu_{22} = 4000 - 80 = 3920$$

Statistical Significance: $\chi 2$ Test

$$
\begin{aligned}
s &= \frac{(90 - 0.6 \times 0.017 \times 10000)^2}{0.6 \times 0.017 \times 10000} + \frac{(80 - 0.4 \times 0.017 \times 10000)^2}{0.4 \times 0.017 \times 10000} \\
&= \frac{(5910 - 0.6 \times 0.983 \times 10000)^2}{0.6 \times 0.983 \times 10000} + \frac{(3920 - 0.4 \times 0.983 \times 10000)^2}{0.4 \times 0.983 \times 10000}
\end{aligned}
$$

Exercise: Calculate p-value of s using `scipy.stats.chi2.cdf` with `df=1`

Statistical Significance: Fisher's Test

Version	n conversions	n sessions - n conversions	total
A	2	18	20
B	5	11	16
Total	7	29	36

Statistical Significance: Fisher's Test

Version	n conversions	n sessions - n conversions	total
A	2	18	20
B	5	11	16
Total	7	29	36

From a population of size $N = 36$ with $K = 7$ conversions we draw a sample of size $n = 20$. What is the distribution of the number of conversion in that sample?

Statistical Significance: Fisher's Test

Version	n conversions	n sessions - n conversions	total
A	2	18	20
B	5	11	16
Total	7	29	36

From a population of size $N = 36$ with $K = 7$ conversions we draw a sample of size $n = 20$. What is the distribution of the number of conversion in that sample?

$x \sim$ Hypergeometric(N=36, K=7, n=20), $P(x = 2) =?$

What is the probability to observe values following the same distribution and as extreme as described $x = 2$?

Statistical Significance: Fishers test

Fisher's exact test: sum of probabilities of over all the tables that yield the observed marginal counts and values of x as extreme as above:

Version	**n conversions**	**n sessions - n conversions**	**total**
A	x	*	20
B	*	*	*
Total	7	29	36

Answer: $P\begin{pmatrix} x=0 & 8 \\ 7 & 21 \end{pmatrix} + P\begin{pmatrix} x=1 & 10 \\ 6 & 19 \end{pmatrix} + P\begin{pmatrix} x=2 & 5 \\ 18 & 11 \end{pmatrix}$

Exercise: use `scipy.stats.fisher_exact` to calculate the corresponding p-value

Limitations: Exact answer to a wrong question?

Statistical Significance: Fishers test

Fisher's exact test: sum of probabilities of over all the tables that yield the observed marginal counts and values of x as extreme as above:

Version	n conversions	n sessions - n conversions	total
A	x	*	20
B	*	*	*
Total	7	29	36

Answer: $P\begin{pmatrix} x=0 & 8 \\ 7 & 21 \end{pmatrix} + P\begin{pmatrix} x=1 & 10 \\ 6 & 19 \end{pmatrix} + P\begin{pmatrix} x=2 & 5 \\ 18 & 11 \end{pmatrix}$

Exercise: use `scipy.stats.fisher_exact` to calculate the corresponding p-value

Limitations: Exact answer to a wrong question? (*The total number of successes in the population is assumed to be fixed*...)

Confidence Interval

Usecase: Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. Construct a 95% confidence interval for the population mean time to complete the tax forms.

Confidence Interval

Definition

Let's say we have a parameter θ of the population (average number of time to complete a tax form). And we have a procedure that produces an estimate $\hat{\theta}$ of this parameter on a sample from the population (sample mean).

Since we sample in a randomized way $\implies$ $\hat{\theta}$ is a random variable

$$P(-\alpha_1 \leq \hat{\theta} - \theta \leq \alpha_2) = 0.95 \implies P(\hat{\theta} - \alpha_2 \leq \theta \leq \hat{\theta} - \alpha_1) = 0.95$$

Confidence Interval: $(\hat{\theta} - \alpha_2,\ \hat{\theta} - \alpha_1)$

Confidence Interval

Numbers α_1, α_2 are chosen in such a way that the confidence interval is symmetric.

Definition

Quantiles: A number α such that $P(X <= \alpha) = p$ is called p-quantile of X.

we find α_1, α_2 s.t. $P(\alpha_1 \leq \hat{\theta} - \theta \leq \alpha_2) = 0.95$

$\implies$ α_1 - 0.025 quantile, α_2 - 0.975 quantile

of the random variable $\hat{\theta} - \theta$

Confidence Interval

Important: Estimate $\hat{\theta}$ is a random variable $\implies$ confidence interval is a pair of random variables!

Correct interpretation: There is a 95% probability that the confidence interval calculated for some future value of the estimate $\hat{\theta}$ will contain the true value of the population parameter.

Wrong interpretation: Let's say we obtained an estimate $\hat{\theta} = 7$ of the population parameter and calculated the corresponding confidence interval $(7 - \alpha_2, 7 - \alpha_1)$. We CANNOT say that there is 95% probability that the true parameter lies in this confidence interval!

Confidence Interval

Important: Estimate $\hat{\theta}$ is a random variable $\implies$ confidence interval is a pair of random variables!

Correct interpretation: There is a 95% probability that the confidence interval calculated for some future value of the estimate $\hat{\theta}$ will contain the true value of the population parameter.

Wrong interpretation: Let's say we obtained an estimate $\hat{\theta} = 7$ of the population parameter and calculated the corresponding confidence interval $(7 - \alpha_2, 7 - \alpha_1)$. We CANNOT say that there is 95% probability that the true parameter lies in this confidence interval!

Difference in statementes: We can talk about $P(X <= 5)$ for a random variable X. But if we have an outcome $X = 7$ of X we cannot talk about the probability that $7 <= 5$!

Confidence Interval

Let $\bar{X} = \sum_{i=1}^{100} X_i$ be the sample mean time. By CLT,

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

Find a value z s.t.

$$P(-z \leq Z \leq z) = 0.95 \quad \Longrightarrow \quad z = q_{0.975} = 1.96$$

Then,

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \\ &= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Confidence Interval

Let $\bar{X} = \sum_{i=1}^{100} X_i$ be the sample mean time. By CLT,

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

Find a value z s.t.

$$P(-z \leq Z \leq z) = 0.95 \quad \Longrightarrow \quad z = q_{0.975} = 1.96$$

Then,

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \\ &= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

For $\bar{X} = 23.6$, the interval is equal to $(22.23, 24.97)$ (`scipy.stats.norm.interval`).

Confidence Interval for unknown distributions

Bootstrap confidence interval Let's say you have a sample of the random variable X: $x_1, \ldots, x_n$ and you can estimate a statistic $\hat{\theta}$ of the parameter θ from this sample (example $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_n$)

- From your sample $x_1, \ldots, x_n$ generate m bootstrap samples of size n (draw with replacement): $(x^*_{11}, \ldots, x^*_{1n}), \ldots, (x^*_{m1}, \ldots, x^*_{mn})$
- Calculate m statics $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_m$ from $(x^*_{11}, \ldots, x^*_{1n}), \ldots, (x^*_{m1}, \ldots, x^*_{mn})$ the same way you calculated $\hat{\theta}$ from $x_1, \ldots, x_n$

Bootstrap Confidence Interval

- θ_i^* approximates $\hat{\theta}$ in the same way as $\hat{\theta}$ approximates θ
- Even if $\hat{\theta}$ is far from θ, the difference $\delta_i^* = \hat{\theta}_i^* - \hat{\theta}$ is close to the difference $\delta = \hat{\theta} - \theta$
- Estimate from the data the 0.025 and 0.975 quatiles $q_{0.025}^*$ and $q_{0.975}^*$ of δ^*
- The approximation of the confidence interval is then given by $(\hat{\theta} - q_{0.025}^*, \hat{\theta} - q_{0.975}^*)$

Bootstrap Confidence Interval

To estimate the quatile $q_{0.025}$ of δ^*:

- calculate $\delta^*_1, \ldots, \delta^*_m$
- order these values from smallest to highest
- calculate the index $k = round(m * 0.025)$
- $q_{0.025} = \delta^*_k$

References

- Grossmann, W., & Rinderle-Ma, S. (2015). Fundamentals of Business intelligence. Springer.
- https://cims.nyu.edu/∼cfgranda/pages/stuff/probability_stats_for_DS.pdf
- https://towardsdatascience.com/the-art-of-a-b-testing-5a10c9bb70a4
- https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24.pdf