MTH102 Engineering Mathematics II

Lesson 10: Limit theorems

Term: 2024

Outline

Outline

Motivations

- Flip a fair coin N times and count the number of heads as N_H. From experience, the larger N is, the closer the ratio $\frac{N_H}{N}$ is to $\frac{1}{2}$. Why?
- In physical measurement experiments, we take repeated measurements and use the average to estimate the exact value. Is it a good idea?
- It is often assumed that the observations of a large sample size are normally distributed. Is it a good hypothesis?

Markov's inequality

Proposition

If X is a random variable that takes only nonnegative values, then for any value $a > 0$,

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Chebyshev's inequality

Proposition

If X is a random variable with mean μ and variance σ^2, then for any value $k > 0$,

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

Or equivalently,

$$P(|X - \mu| < k) \geq 1 - \frac{\sigma^2}{k^2}.$$

Example 1

Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.

(a) What can be said about the probability that this week's production will exceed 75?

(b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

Outline

Sample mean

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables, each having mean μ and variance σ^2. Then the random variable

$$\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$$

is called the *sample mean* of $X_1, X_2, \ldots, X_n$, and

$$E(\bar{X}) = \mu, \;\; Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Note that as n increases, the variance of $\bar{X}$ decreases. Roughly speaking, the observations on $\bar{X}$ will be closer to μ as n increases.

Central limit theorem

Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \left(= \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \right)$$

tends to the standard normal distribution $N(0,1)$ as $n \to \infty$. That is, for any $a \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq a \right) = \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} \, dx.$$

Example 3

A teacher has 25 exam papers that will be graded in sequence. The times required to grade the exam papers are independent, with a common distribution that has mean 20 minutes and standard deviation 4 minutes. Approximate the grading work will be done in 8 hours.

Solution. For $i = 1, 2, \ldots, 25$, let X_i be the time in minutes to grade the i-th exam paper. Then $\mu = E(X_i) = 20$ and $\sigma = 4$.

$$\begin{aligned} P\left(\sum_{i=1}^{25} X_i \leq 480\right) &= P\left(\frac{\sum_{i=1}^{25} X_i - 25 \times 20}{4\sqrt{25}} \leq \frac{480 - 500}{20}\right) \\ &= \Phi(-1) \\ &= 0.1587. \end{aligned}$$

Example 4

A die is continually rolled until the total sum of all rolls exceeds 350. Approximate the probability that at least 105 rolls are necessary.

Solution. For $i = 1, 2, \ldots, 105$, let X_i be the number of the i-th roll. Then

$$\mu = \frac{7}{2}, \ \sigma^2 = \frac{35}{12}.$$

$$\begin{aligned} P\left(\sum_{i=1}^{105} X_i \geq 350\right) &= P\left(\frac{\sum_{i=1}^{105} X_i - 105\mu}{\sigma\sqrt{105}} \geq -1\right) \\ &= 1 - \Phi(-1) \\ &= 0.8413. \end{aligned}$$

Exercise

A person has 100 light bulbs whose lifetimes are independent exponentials with mean 5 hours. If the bulbs are used one at a time, with a failed bulb being replaced immediately by a new one, approximate the probability that there is still a working bulb after 525 hours.

Outline

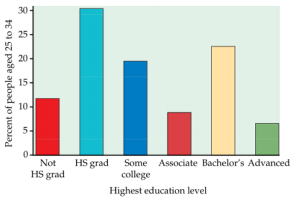
Sample

- We are interested in a random variable X, but the distribution of X is unknown or the information is incomplete.
- We perform the random experiments n times, obtaining n observed values of the random variable X: $x_1, x_2, \ldots, x_n$.
- The collection of data $x_1, x_2, \ldots, x_n$ is referred to as a **sample**.
- From the sample data, we investigate some key information about the distribution of X.

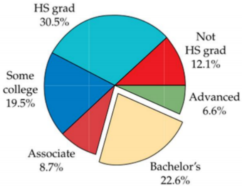
Visualization of data

Data on the educational level of a certain population.

Bar graph



Pie Charts



Various plot options: histograms, stem and leaf plot, scatter plot, boxplot...

Descriptive statistics

From the collection of data $x_1, x_2, \dots, x_n$, we want to describe some essential features of the distribution by a summary of the data $g(x_1, x_2, \dots, x_n)$ which is called **statistics**. In particular, we are interested in the following.

- **Central tendency**: a single numerical value considered as "the most typical of data".
- **Spread**: how much the data are different from the central value.

Descriptive statistics: central tendency

From the collection of data $x_1, x_2, \ldots, x_n$, we want to describe the central tendency which is a single numerical value considered as "the most typical of data".

- **Mean** (or **sample mean**):

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

- **Median**: the center point m in the set of ordered data. Given a data sample in ascending order: $x_1, x_2, \ldots, x_n$.

$$m = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd,} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even.} \end{cases}$$

- **Mode**: the most frequently occurring number.

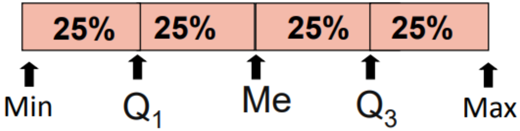
Central tendency: Example 6

- The sample data is $1, 2, 3, 4, 5$, the mean is 3, the median is 3, and the mode is every number.
- The sample data is $1, 2, 3, 4, 5, 5$, the mean is $\frac{10}{3}$, the median is 3.5, and the mode is 5.

Descriptive statistics: quartiles

Given a data sample in ascending order: $x_1, x_2, \ldots, x_n$.

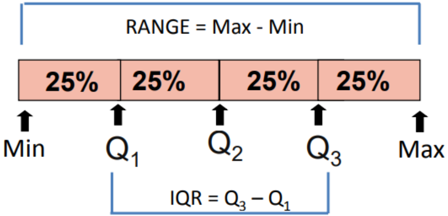
- **Quartiles** split the data into 4 quarters.
- The middle one Q_2 is the median.
- If n is odd, Q_1 is the median of $x_1, x_2, \ldots, x_{\frac{n+1}{2}-1}$ and Q_3 is the median of $x_{\frac{n+1}{2}+1}, \ldots, x_n$.
- If n is even, Q_1 is the median of $x_1, x_2, \ldots, x_{\frac{n}{2}}$ and Q_3 is the median of $x_{\frac{n}{2}+1}, \ldots, x_n$.



Descriptive statistics: spread

From the collection of data $x_1, x_2, \ldots, x_n$, we want to describe the variability of the data.

- **Range**: the difference between the maximal and minimal numbers.
- **Interquartile range IQR**: the difference between Q_3 and Q_1.



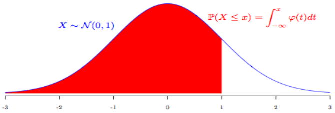
- **Variance** (or **Sample variance**):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Example 7

Consider a sample of 16 data: $1, 3, 4, 5, 5, 6, 8, 10, 15, 15, 16, 18, 18, 20, 20, 26$.

- Mean: 11.875.
- Median: $\frac{x_8+x_9}{2} = \frac{10+15}{2} = 12.5$.
- Q_1 is the median of $x_1, \ldots, x_8$: $\frac{5+5}{2} = 5$.
- Q_3 is the median of $x_9, \ldots, x_{16}$: $\frac{18+18}{2} = 18$.
- Range: $26 - 1 = 25$.
- IQR: $Q_3 - Q_1 = 18 - 5 = 13$.
- $s^2 = 56.65$.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Figure: cdf for standard normal r.v.