STAT 516

Lecture 8: Normal distribution

Prof. Michael Levine

April 3, 2020

Introduction

- Most empirical data that seem to be unimodal and not strongly skewed are commonly modeled using the normal distribution
- When a new methodology is presented, it is typically tested on the normal distribution first
- The best-known procedures in statistics have their exact inferential optimality properties when the data come from the normal distribution

Definition

- $X \sim N(\mu, \sigma^2)$ when its pdf is
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
for any $-\infty < x < \infty$
- In this definition, μ can be any real number and $\sigma > 0$.
- The case $X \sim N(0, 1)$ is called a standard normal random variable

Definition

- The density of the standard normal random variable is denoted as $\phi(x)$ and is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

  for any $-\infty < x < \infty$
- $\phi(x)$ is symmetric and unimodal about zero. The general $N(\mu, \sigma^2)$ is symmetric and unimodal around μ

Definition

- By definition, the CDF of the standard normal distribution is

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)\, dz$$

- Due to the symmetry of the standard normal distribution around zero

$$\Phi(-x) = 1 - \Phi(x)$$

- The change of μ results in the shift of the distribution to the new center
- The increase of σ^2 results in the new distribution being more spread out

Standard Normal CDF at Selected Values

x	$\Phi(x)$
-4	0.0003
-3	0.00135
-2	0.02275
-1	0.15866
0	0.5
1	0.84134
2	0.97725
3	0.99865
4	0.99997

Basic properties

- If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$; if $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$
- If $X \sim N(\mu, \sigma^2)$, then

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

  In particular, $P(X \leq \mu) = P(Z \leq 0) = 0.5$ i.e. μ is the median of X
- Every moment of any normal distribution exists; for any k, $E[(X-\mu)^{2k+1}] = 0$

Basic properties

- If $Z \sim N(0, 1)$, then
$$E(Z^{2k}) = \frac{(2k)!}{2^k k!}$$
for any $k \geq 1$
- The MGF of $N(\mu, \sigma^2)$ exists for all real t and is
$$\psi(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}$$

Corollary

- Let $X \sim N(\mu, \sigma^2)$ and $0 < \alpha < 1$.
- Let $Z \sim N(0, 1)$ and denote x_α the $(1-\alpha)$th quantile of X
- Finally, let z_α is the $(1-\alpha)$th percentile of Z. Then,

$$x_\alpha = \mu + \sigma z_\alpha$$

Basic Example I

- By using a standard normal CDF table, we can easily find 75th, 90th, 97.5th, 99th, and 99.5th percentiles of the standard normal distribution

α	$1-\alpha$	z_α
0.25	0.75	0.675
0.1	0.9	1.282
0.05	0.95	1.645
0.025	0.975	1.960
0.01	0.99	2.326
0.005	0.995	2.576

Basic Example II

- The age of the subscribers to a newspaper has a normal distribution with mean 50 years and standard deviation 5 years. Compare the percentage of subscribers who are less than 40 years old and the percentage who are between 40 and 60 years old.
- $X \sim N(\mu, \sigma^2)$ with $\mu = 50$ and $\sigma = 5$ is the age of a subscriber. Then,

$$P(X < 40) = \Phi\left(\frac{40 - 50}{5}\right) = \Phi(-2) = 0.02275$$

and

$$P(40 \leq X \leq 60) = P(X \leq 60) - P(X \leq 40)$$
$$= \Phi(2) - \Phi(-2) = 0.9545$$

Example I

- Let X denote the length of time (in minutes) an auto battery will continue to crank an engine. Assume that $X \sim N(10, 4)$.
- What is the probability that the battery will crank the engine longer than $10 + x$ minutes given that it is still cranking in 10 minutes?
- $$P(X > 10 + x | X > 10) = \frac{P(X > 10 + x)}{P(X > 10)} = \frac{P(Z > x/2)}{1/2}$$
  $$= 2\left[1 - \Phi\left(\frac{x}{2}\right)\right]$$
- Note that the resulting function is decreasing in x.
- This is different from the exponential distribution with the same mean $\mu = 10$

Example II

- Let the thermostat be set at d degrees Celsius.
- The actual temperature of a certain room is $N(d, \sigma^2)$ with $\sigma = 0.5$
- If the thermostat is set at 75 degrees, what is the probability that the actual temperature is below 74 degrees?
- $$P(X < 74) = P(Z < (74 - 75)/0.5) = P(Z < -2) = 0.02275$$

Example II

- What is the lowest setting of the thermostat that will maintain a temperature of at least 72 degrees with probability of 0.99?
- We need to find the value of d such that $P(X \geq 72) = 0.99$, or equiv. $P(X < 72) = 0.01$
- Note that $P(Z < -2.36) = 0.01$ (e.g. see the normal distribution table or use the software)
- Thus, need to find d such that $d + \sigma(-2.326) = 72$ which results in $d = 73.16$ degrees

Sums of independent normal variables

- Let $X_1, \dots, X_n$ for $n \geq 2$ be independent random variables $X_i \sim N(\mu_i, \sigma_i^2)$.
- Also, let $S_n = \sum_{i=1}^n X_i$.
- Then,
$$S_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$
- A sum of any number of independent normal random variables is exactly normally distributed
- Note that a more general statement is also true: for any set of constants $a_1, \dots, a_n$
$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

Proof

- The mgf of S_n is

$$\psi_{S_n}(t) = E(e^{tS_n}) = E(e^{tX_1} \cdots e^{tX_n}) = \prod_{i=1}^{n} E(e^{tX_i})$$

$$= \prod_{i=1}^{n} e^{t\mu_i + t^2\sigma_i^2/2} = e^{t(\sum_{i=1}^{n} \mu_i) + (t^2/2)(\sum_{i=1}^{n} \sigma_i^2)}$$

- The last expression is the mgf of $N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$

Corollary

- Suppose X_i, $1 \leq i \leq n$ are independent and each is distributed as $N(\mu, \sigma^2)$.
- Then, $\bar{X} = \frac{S_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- Thus, the distribution of $\bar{X}$ becomes more concentrated around the true mean μ as the sample size grows.
- Therefore, $\bar{X}$ becomes better and better as an estimator of μ.

Example I

- Suppose $X \sim N(-1, 4)$, $Y \sim N(1, 5)$ and they are independent.
- What is the CDF of $X + Y$ and $X - Y$?
- First, $X + Y \sim N(0, 9)$ and $X - Y \sim N(-2, 9)$
- Therefore, $P(X + Y \leq x) = \Phi\left(\frac{x}{3}\right)$ and $P(X - Y \leq x) = \Phi\left(\frac{x+2}{3}\right)$

Example II

- Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are independent (they are iid)
- Therefore, $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$ and

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

- Thus, with probability 95% for any n we have that the true mean μ is between $\bar{X} - 1.96\sigma/\sqrt{n}$ and $\bar{X} + 1.96\sigma/\sqrt{n}$
- We obtained a simple 95% confidence interval for μ with the margin of error $1.96\sigma/\sqrt{n}$.