The Normal Distributions

Introduction to Normal Distributions

- Normal distributions are often referred to as "bell-shaped curves."
- They are characterized by their symmetric, bell-shaped density functions.

Density and Properties

- The density of a normal distribution is given by the equation:

$$f_W(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}, -\infty < t < \infty \tag{1}$$

- It's a two-parameter family, indexed by μ (mean) and σ (standard deviation).
- The notation $N(\mu, \sigma^2)$ is used, where σ^2 is the variance.

Notation and Implications

- The notation $X \sim N(\mu, \sigma^2)$ indicates that the random variable X follows the normal distribution $N(\mu, \sigma^2)$.
- Important Note:
  - Saying “X has a $N(\mu, \sigma^2)$ distribution” implies more than just the mean and variance.
  - It also indicates that X has a bell-shaped density within the normal distribution family.

Closure Under Affine Transformation

Affine Transformation of Normal Distributions

- The normal distribution family is closed under affine transformations.
- This means that if $X \sim N(\mu, \sigma^2)$, and $Y = cX + d$, then Y also follows a normal distribution.

Mathematical Formulation:

If

$$X \sim N(\mu, \sigma^2) \tag{2}$$

and we set

$$Y = cX + d \tag{3}$$

then

$$Y \sim N(c\mu + d, c^2\sigma^2) \tag{4}$$

Practical Example

- Consider X as the height of a AU student in inches.
- If Y is the height in centimeters, then $c = 2.54$ and $d = 0$.
- This implies that the histogram of Y will also be bell-shaped.

Deeper Understanding

- It's more than just Y having mean $c\mu + d$ and variance $c^2\sigma^2$.
- The key point is that Y remains a member of the normal family with its density still defined by the normal distribution formula.

Derivation

- Assume $c > 0$. Then the derivation follows several steps involving the distribution functions F_X and F_Y, and their derivatives.
- The final expression confirms that Y has a $N(c\mu + d, c^2\sigma^2)$ distribution.

By definition:

$$F_Y(y) = P(Y \leq y)$$

Substituting $Y = cX + d$ into the inequality:

$$P(cX + d \leq y) = P\left(X \leq \frac{y-d}{c}\right) = F_X\left(\frac{y-d}{c}\right)$$

Derivation

Substitute the expression for $F_X(x)$ into the equation:

$$F_Y(y) = \Phi\left(\frac{\frac{y-d}{c} - \mu}{\sigma}\right) = \Phi\left(\frac{y - d - c\mu}{c\sigma}\right)$$

Then, the pdf of Y is

$$f_Y(y) = \frac{d}{dy}\Phi\left(\frac{y - d - c\mu}{c\sigma}\right) = \frac{1}{c\sigma}\phi\left(\frac{y - d - c\mu}{c\sigma}\right)$$

where ϕ is the pdf of $N(\mu, \sigma^2$. Thus,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}c\sigma}\ e^{-\frac{1}{2}\left(\frac{y-(c\mu+d)}{c\sigma}\right)^2}$$

which is the pdf of $N(c\mu + d, c^2\sigma^2)$.

Closure Under Independent Summation in Normal Distributions

Independent Summation

- If X and Y are independent random variables, each normally distributed, then their sum $S = X + Y$ is also normally distributed.
- This property is unique to the normal distribution and not observed in most other distributions.

Comparative Example

- For contrast, if X and Y each have a uniform distribution U(0,1), the distribution of their sum S is triangular, not uniform.
- This difference highlights the unique nature of the normal distribution.

General Case

- For any constants c and d, if X and Y are independent and normally distributed, then $cX + dY$ will also have a normal distribution.
- More generally, for constants $a_1, \ldots, a_k$ and independent normal random variables $X_1, \ldots, X_k$:

$$Y = a_1 X_1 + \ldots + a_k X_k \implies Y \sim N\left(\sum_{i=1}^{k} a_i \mu_i, \sum_{i=1}^{k} a_i^2 \sigma_i^2\right) \quad (5)$$

Lack of Intuition

- The property that $X + Y$ is normally distributed when X and Y are independent and normally distributed is counterintuitive.
- There's no straightforward intuitive explanation for why the sum of two normal distributions remains normally distributed.

R Functions for Normal Distributions

```
dnorm(x, mean = 0, sd = 1)
pnorm(q, mean = 0, sd = 1)
qnorm(p, mean = 0, sd = 1)
rnorm(n, mean = 0, sd = 1)
```

- These functions are used for different purposes related to the normal distribution in R.
- **mean** and **sd** represent the mean and standard deviation of the distribution.

The Standard Normal Distribution

Definition of Standard Normal Distribution

Definition
If $Z \sim N(0, 1)$, then the random variable Z has a *standard normal distribution*.

- This is a special case where the mean (μ) is 0 and the standard deviation (σ) is 1.

Transforming to Standard Normal

- For any normal random variable X with $X \sim N(\mu, \sigma^2)$:

$$Z = \frac{X - \mu}{\sigma} \tag{6}$$

- This transformation results in:

$$Z \sim N(0, 1) \tag{7}$$

Derivation and Properties

- Start with $Z = \frac{X-\mu}{\sigma}$ and rewrite it as $Z = \frac{1}{\sigma} \cdot X + \left(\frac{-\mu}{\sigma}\right)$.
- For any random variable U and constants c and d, $E(cU + d) = cEU + d$.
- Thus, $EZ = \frac{1}{\sigma}EX - \frac{\mu}{\sigma} = 0$.
- Using the properties of variance, we find that Z has variance 1.
- Due to the closure under affine transformations, Z retains a normal distribution.

Cumulative Distribution Function (CDF)

- The cumulative distribution function (CDF) of the standard normal distribution is traditionally denoted by Φ.

Evaluating Normal cdfs

- The normal distribution function does not have a closed-form definite integral.
- Traditionally, cdf values for the standard normal distribution (N(0,1)) are used for approximating probabilities.
- A table for the N(0,1) cdf is often included in statistics textbooks.

One Table for the Entire Normal Family

- Though there are infinitely many distributions in the normal family, one table for N(0,1) suffices.
- Example: For $X \sim N(10, 2.5^2)$ and calculating $P(X < 12)$:

$$P(X < 12) = P\left(\frac{X - 10}{2.5} < \frac{12 - 10}{2.5}\right) = P(Z < 0.8) \approx .79 \tag{8}$$

- Here, Z is the standard normal variable, and its probability can be found using the N(0,1) table.

Using R for Normal Distributions

- The R statistical package provides functions for working with normal distributions.
- **pnorm()** for the normal cdf:

```
pnorm(q, mean = 0, sd = 1)
```

- **rnorm()** for simulating normal random variables:

```
rnorm(n, mean = 0, sd = 1)
```

- **dnorm()** and **qnorm()** for density and quantile functions.

Network Intrusion: The Scenario

- Jill's remote logins involve reading/writing disk sectors.
- The number of sectors (X) is approximately normally distributed with mean 500 and standard deviation 15.
- Modeling note: The number of sectors is discrete, but can be approximated as a continuous normal distribution.

Analyzing Suspicious Activity

- Scenario: A login (possibly Jill's) reads/writes 535 sectors.
- Question: Should this be considered suspicious?
- Approach: Calculate $P(X \geq 535)$.

Probability Calculation

To find $P(X \geq 535)$:

- Transform to standard normal: $Z = \frac{X-500}{15}$.
- Calculate:

$$P(X \geq 535) = P\left(\frac{X-500}{15} \geq \frac{535-500}{15}\right) = P\left(Z \geq \frac{35}{15}\right) = 1 - \Phi(2.33)$$

Using R for Probability Calculation:

In R, use **pnorm()** to compute this probability.

```
> 1 - pnorm(535, 500, 15)
[1] 0.009815329
```

- Probability of ≈ 0.01 makes the activity suspicious.
- Further investigation recommended based on this probability.

Analyzing Two Suspicious Logins

- Scenario: Two logins to Jill's account, with $X + Y = 1088$ sectors accessed.
- Assumption: X and Y are independent.
- S = X + Y is normally distributed with mean 1000 and variance 2×15^2.
- Calculate $P(X + Y > 1088)$.

Probability Calculation for Two Logins:

```
> 1 - pnorm(1088, 1000, sqrt(450))
[1] 1.674329e-05
```

- This returns a very small probability, indicating high suspicion.
- Such rare events warrant further investigation.

The Central Limit Theorem

Introduction to the Central Limit Theorem

- The Central Limit Theorem (CLT) states that a random variable, which is a sum of many components, will have an approximate normal distribution.
- Examples include human weights and raw SAT test scores.

The Basic Central Limit Theorem:

Theorem
Suppose $X_1, X_2, \ldots$ are independent random variables, all having the same distribution with mean m and variance v^2. Form the new random variable $T = X_1 + \ldots + X_n$. Then for large n, the distribution of T is approximately normal with mean nm and variance nv^2.

Requirements and Approximation

- Requirements for the CLT:
  - Summands must be independent and identically distributed.
  - Distribution of each summand should have a finite mean and variance.
- The larger n is, the better the approximation.
- Typically, $n = 20$ or even $n = 10$ is sufficient for a good approximation.

Implications of the CLT

- The CLT explains why many real-world phenomena follow a normal distribution.
- It is a fundamental concept in statistics and probability, underlying many statistical methods and analyses.

Example 1: Sum of Uniform Distributions

- Consider $W = U_1 + \ldots + U_{50}$ with U_i being i.i.d. and uniformly distributed on (0,1).
- Goal: Approximate $P(W < 23.4)$.
- W has an approximate normal distribution with mean 50×0.5 and variance $50 \times \frac{1}{12}$.
- R Evaluation:

```
> pnorm(23.4, 25, sqrt(50/12))
[1] 0.216568
```

Example 2: Bug Counts

- Bugs per 1,000 lines of code follow a Poisson distribution with mean 5.2 where 1000 lines of code constitutes a section.
- Find the probability of more than 106 bugs in 20 sections.
- Assumption: Sections act independently.
- R Evaluation:

```
> 1 - pnorm(106, 20*5.2, sqrt(20*5.2))
[1] 0.4222596
```

Example 3: Coin Tosses

- Binomial distributions with large *n* are approximately normally distributed (CLT).
- Example: Approximate probability of more than 12 heads in 20 tosses.
- R Evaluation:

```
> 1 - pnorm(12, 10, sqrt(5))
[1] 0.1855467
```

- Exact answer: 0.132, but approximation gives 0.186.
- Improved accuracy with correction for continuity:

```
> 1 - pnorm(12.5, 10, sqrt(5))
[1] 0.1317762
```

- This correction brings the approximation closer to the exact answer.

The Importance of Normal Distribution in Statistical Modeling

Real World and Normal Distribution

- No real-world random variables are exactly normally distributed.
- Real-world variables don't have continuous distributions and are bounded, unlike normal distributions which extend from $-\infty$ to ∞.

Approximate Normal Distributions in Nature:

- Many natural phenomena have approximate normal distributions.
- This approximation plays a key role in statistical methods and analysis.
- Classical statistical procedures often assume sampling from approximately normal populations.

The Central Limit Theorem and Normal Distribution

- The Central Limit Theorem (CLT) implies that quantities used for statistical estimation are approximately normal, even if the underlying data are not.
- This is significant in cases where the data itself might not be normally distributed.
- Example: The gamma distribution, or Erlang distribution, becomes approximately normal for large values due to the CLT.

Conclusion

- The normal distribution model is a useful approximation for real-world data analysis.
- While simplistic, this example illustrates the fundamental concepts in intrusion detection analysis.
- Despite theoretical limitations, the normal distribution is a powerful and versatile tool in statistics.
- Its significance is enhanced by the CLT, making it relevant in a wide range of practical applications.

The Chi-Squared Family of Distributions

Chi-Squared Distribution: Definition

- Defined as the distribution of $Y = Z_1^2 + \ldots + Z_k^2$, where $Z_1, Z_2, \ldots, Z_k$ are independent N(0,1) random variables.
- Noted as χ_k^2 and called chi-squared with k degrees of freedom.
- A one-parameter family of distributions frequently used in statistical applications.

Mean and Variance of Chi-Squared Distribution

- The mean EY of chi-squared distribution is k.
- Derived as $EY = E(Z_1^2 + \ldots + Z_k^2) = kE(Z_1^2)$ and $E(Z_1^2) = Var(Z_1) + [E(Z_1)]^2 = 1$. Hence $EY = \underbrace{(1 + \ldots + 1)}_{k \text{ many}} = k$
- The variance $Var(Y)$ is $2k$.

Relation to Gamma Family

- Chi-squared is a special case of the gamma family.
- Corresponds to gamma distribution with $r = k/2$ and $\lambda = 0.5$.

R Functions for Chi-Squared Distribution

- **dchisq()**, **pchisq()**, **qchisq()**, **rchisq()** for density, CDF, quantile function, and random number generation.
- Example: To get the density value $f_X(5.2)$ for a chi-squared random variable with 3 degrees of freedom:

```
> dchisq(5.2, 3)
[1] 0.06756878
```

Chi-Squared Distribution: Pin Placement Error Example

- Machine places a pin in the middle of a disk-shaped object.
- X and Y: placement errors in horizontal and vertical directions, respectively.
- X and Y are independent, normally distributed with mean 0 and variance 0.04.
- Goal: Find $P(W > 0.6)$ where W is the distance from true center to pin placement.

Transforming the Problem

- Distance W is the square root of a sum of squares: $W^2 = X^2 + Y^2$.
- Transform the problem: $P(W > 0.6) = P(W^2 > 0.36)$.
- Convert to a chi-squared problem: $P[(X/0.2)^2 + (Y/0.2)^2 > 0.36/0.2^2] = P[\chi^2_2 > 9]$.

R Evaluation:

- The problem now fits the chi-squared distribution with 2 degrees of freedom.
- R code to evaluate $P(W > 0.6)$:

```
> 1 - pchisq(0.36/0.04, 2)
 [1] 0.01110900
```

- This gives the probability of the pin placement error being greater than 0.6.

Generating Normal Random Numbers

- Normal random number generators like **rnorm()** use the relationship between normal and exponential distributions.
- Define $W = Z_1^2 + Z_2^2$ with Z_1 and Z_2 as independent N(0,1) random variables.
- W follows a chi-squared distribution with 2 degrees of freedom, equivalent to an exponential distribution with $\lambda = 0.5$.

Generating N(0,1) Random Variates

- Using the transformation $\theta = \tan^{-1}(Z_2/Z_1)$, where θ is uniformly distributed on $(0, 2\pi)$.
- Express Z_1 and Z_2 in terms of W and θ:
  $Z_1 = \sqrt{W}\cos(\theta), \quad Z_2 = \sqrt{W}\sin(\theta).$
- R code to generate a pair of independent N(0,1) random variates:

```
genn01 <- function () {
theta <- runif (1 ,0 ,2*pi)
w <- rexp (1 ,0.5)
sw <- sqrt (w)
c(sw*cos( theta ),sw*sin( theta ))
}
```

Importance of Chi-Squared in Modeling

- Chi-squared distribution is widely used in statistical methods.
- It often arises in sums of squared normal random variables.
- The term "degrees of freedom" in this context will be explained in later chapters on statistics.

Relation to Gamma Family

- The chi-square distribution with d degrees of freedom is a gamma distribution.
- Corresponds to a gamma distribution with $r = d/2$ and $\lambda = 0.5$.

The Multivariate Normal Family

Introduction to Multivariate Normal Family

- Generalization of the normal family to multiple dimensions.
- Parameterized by a vector mean and a covariance matrix.

Bivariate Normal Distribution:

- Bivariate normal distribution for joint distribution of X_1 and X_2.
- Parameters: means (μ_1, μ_2), standard deviations (σ_1, σ_2), and correlation (ρ).
- Density function is complex, but important for conceptual understanding.

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right)$$

Density of Multivariate Normal Distribution

- For a random vector $X = (X_1, \ldots, X_k)'$ with a k-variate normal distribution:
- Density function:

$$f_X(t) = ce^{-\frac{1}{2}(t-\mu)'\Sigma^{-1}(t-\mu)} \tag{9}$$

- Here c is a constant and Σ is the covariance matrix.

Multivariate Central Limit Theorem

- Sums of random vectors have approximately multivariate normal distributions.

R Functions for Multivariate Normal Distribution

- Density, CDF, and quantiles: **dmvnorm()**, **pmvnorm()**, **qmvnorm()** from the **mvtnorm** library.
- Simulation: **mvrnorm()** from the **MASS** library.