CS 147:
Computer Systems Performance Analysis
Review of Statistics

2015-06-15

CS147

CS 147:
Computer Systems Performance Analysis
Review of Statistics

Introduction to Statistics

- Concentration on applied statistics
- Especially those useful in measurement
- Today's lecture will cover 15 basic concepts
- You should already be familiar with them

Introduction to Statistics

- Concentration on applied statistics
- Especially those useful in measurement
- Today's lecture will cover 15 basic concepts
- You should already be familiar with them

1. Independent Events

- Occurrence of one event doesn’t affect probability of other
- Examples:
  - Coin flips
  - Inputs from separate users
  - “Unrelated”traffic accidents

2015-06-15

CS147
└15 Concepts
  └Independent Events
    └1. Independent Events

1. Independent Events

- Occurrence of one event doesn’t affect probability of other
- Examples:
  - Coin flips
  - Inputs from separate users
  - “Unrelated”traffic accidents

1. Independent Events

- Occurrence of one event doesn't affect probability of other
- Examples:
  - Coin flips
  - Inputs from separate users
  - “Unrelated”traffic accidents
- What about second basketball free throw after the player misses the first?

1. Independent Events

- Occurrence of one event doesn't affect probability of other
- Examples:
  - Coin flips
  - Inputs from separate users
  - “Unrelated”traffic accidents
- What about second basketball free throw after the player misses the first?

2. Random Variable

- Variable that takes values probabilistically
- Variable usually denoted by capital letters, particular values by lowercase
- Examples:
  - Number shown on dice
  - Network delay
  - CS 70 attendance
- What about disk seek time?

2. Random Variable

- Variable that takes values probabilistically
- Variable usually denoted by capital letters, particular values by lowercase
- Examples:
  - Number shown on dice
  - Network delay
  - CS 70 attendance
- What about disk seek time?

3. Cumulative Distribution Function (CDF)

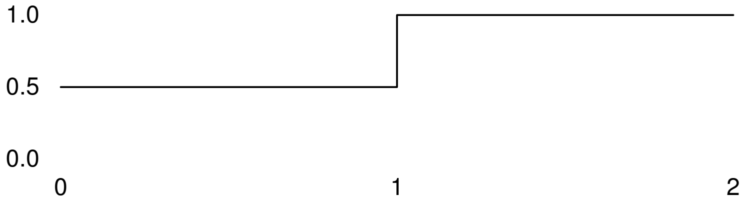
- Maps a value a to probability that the outcome is less than or equal to a:
$$F_x(a) = P(x \leq a)$$
- Valid for discrete and continuous variables
- Monotonically increasing
- Easy to specify, calculate, measure

3. Cumulative Distribution Function (CDF)

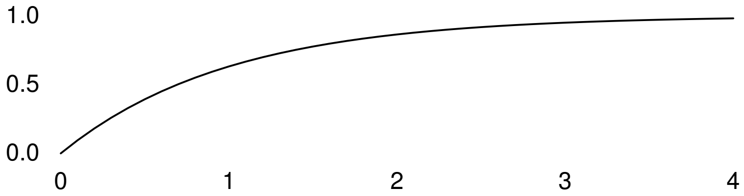
- Maps a value a to probability that the outcome is less than or equal to a:
$$F_x(a) = P(x \leq a)$$
- Valid for discrete and continuous variables
- Monotonically increasing
- Easy to specify, calculate, measure

CDF Examples

- Coin flip (T = 0, H = 1):

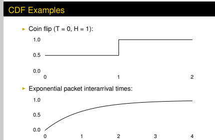


- Exponential packet interarrival times:



2015-06-15

CS147
└15 Concepts
└CDF
└CDF Examples



4. Probability Density Function (pdf)

- Derivative of (continuous) CDF:

$$f(x) = \frac{dF(x)}{dx}$$

- Usable to find probability of a range:

$$\begin{aligned} P(x_1 < x \leq x_2) &= F(x_2) - F(x_1) \\ &= \int_{x_1}^{x_2} f(x)\, dx \end{aligned}$$

4. Probability Density Function (pdf)

- Derivative of (continuous) CDF:

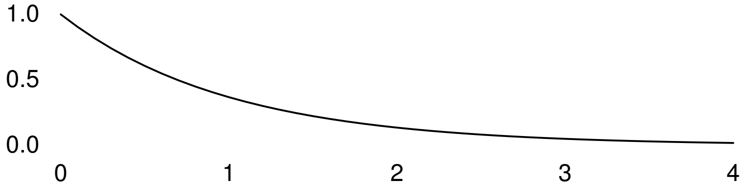
$$f(x) = \frac{dF(x)}{dx}$$

- Usable to find probability of a range:

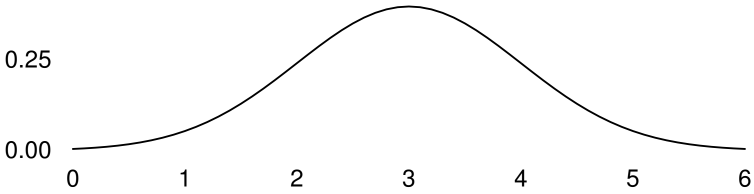
$$\begin{aligned} P(x_1 < x \leq x_2) &= F(x_2) - F(x_1) \\ &= \int_{x_1}^{x_2} f(x)\, dx \end{aligned}$$

Examples of pdf

- Exponential interarrival times:

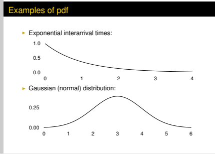


- Gaussian (normal) distribution:



2015-06-15

CS147
└15 Concepts
  └pdf
    └Examples of pdf


Examples of pdf
Exponential interarrival times:
Gaussian (normal) distribution:

5. Probability Mass Function (pmf)

- CDF not differentiable for discrete random variables
- pmf serves as replacement: $f(x_i) = p_i$ where p_i is the probability that x will take on the value x_i:

$$
\begin{aligned}
P(x_1 < x \leq x_2) &= F(x_2) - F(x_1) \\
&= \sum_{x_1 < x \leq x_2} p_i
\end{aligned}
$$

5. Probability Mass Function (pmf)

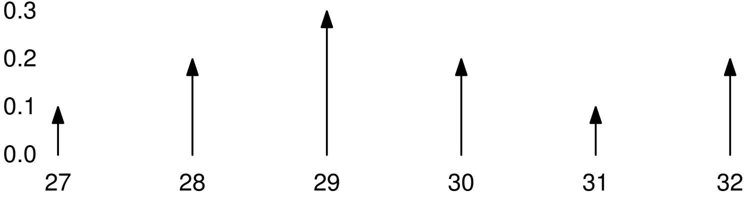
- CDF not differentiable for discrete random variables
- pmf serves as replacement: $f(x_i) = p_i$ where p_i is the probability that x will take on the value x_i:

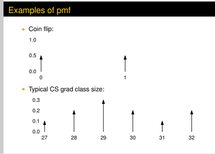
$$
\begin{aligned}
P(x_1 < x \leq x_2) &= F(x_2) - F(x_1) \\
&= \sum_{x_1 < x \leq x_2} p_i
\end{aligned}
$$

Examples of pmf

- Coin flip:

- Typical CS grad class size:





6. Expected Value (Mean)

- Mean:

$$\mu = E(x) = \sum_{i=1}^{n} p_i x_i = \int_{-\infty}^{\infty} x f(x)\, dx$$

- Summation if discrete
- Integration if continuous

6. Expected Value (Mean)
- Mean:
$$\mu = E(x) = \sum_{i=1}^{n} p_i x_i = \int_{-\infty}^{\infty} x f(x)\, dx$$
- Summation if discrete
- Integration if continuous

7. Variance

- Variance:

$$\begin{aligned}\mathrm{Var}(x) = E[(x-\mu)^2] &= \sum_{i=1}^{n} p_i (x_i - \mu)^2 \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx\end{aligned}$$

- Often easier to calculate equivalent $E(x^2) - E(x)^2$
- Usually denoted σ^2; square root σ is called *standard deviation*

7. Variance

- Variance:

$$\begin{aligned}\mathrm{Var}(x) = E[(x-\mu)^2] &= \sum_{i=1}^{n} p_i (x_i - \mu)^2 \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx\end{aligned}$$

- Often easier to calculate equivalent $E(x^2) - E(x)^2$
- Usually denoted σ^2; square root σ is called *standard deviation*

8. Coefficient of Variation (C.O.V. or C.V.)

- Ratio of standard deviation to mean:

$$\text{C.V.} = \frac{\sigma}{\mu}$$

- Indicates how well mean represents the variable

8. Coefficient of Variation (C.O.V. or C.V.)

- Ratio of standard deviation to mean:

$$\text{C.V.} = \frac{\sigma}{\mu}$$

- Indicates how well mean represents the variable

9. Covariance

- Given x, y with means x and y, their covariance is:

$$\begin{aligned}\mathrm{Cov}(x,y) = \sigma^2_{xy} &= E[(x-\mu_x)(y-\mu_y)] \\ &= E(xy) - E(x)E(y)\end{aligned}$$

  - Two typos on p.181 of book
- High covariance implies y departs from mean whenever x does

9. Covariance

- Given x, y with means x and y, their covariance is:

$$\begin{aligned}\mathrm{Cov}(x,y) = \sigma^2_{xy} &= E[(x-\mu_x)(y-\mu_y)] \\ &= E(xy) - E(x)E(y)\end{aligned}$$

  - Two typos on p.181 of book
- High covariance implies y departs from mean whenever x does

Covariance (cont'd)

- For independent variables, $E(xy) = E(x)E(y)$ so $\text{Cov}(x, y) = 0$
- Reverse isn't true: $\text{Cov}(x, y) = 0$ does **NOT** imply independence
- If $y = x$, covariance reduces to variance

Covariance (cont'd)

- For independent variables, $E(xy) = E(x)E(y)$ so $\text{Cov}(x, y) = 0$
- Reverse isn't true: $\text{Cov}(x, y) = 0$ does **NOT** imply independence
- If $y = x$, covariance reduces to variance

10. Correlation Coefficient

- Normalized covariance:

$$\text{Correlation}(x, y) = \rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

- Always lies between -1 and 1
- Correlation of 1 $\Rightarrow x \sim y$, -1 $\Rightarrow x \sim \frac{1}{y}$

10. Correlation Coefficient
- Normalized covariance:
$$\text{Correlation}(x, y) = \rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$
- Always lies between -1 and 1
- Correlation of 1 $\Rightarrow x \sim y$, -1 $\Rightarrow x \sim \frac{1}{y}$

11. Mean and Variance of Sums

- For any random variables,

$$E(a_1 x_1 + \cdots + a_k x_k) = a_1 E(x_1) + \cdots + a_k E(x_k)$$

- For independent variables,

$$\mathrm{Var}(a_1 x_1 + \cdots + a_k x_k) = a_1^2 \mathrm{Var}(x_1) + \cdots + a_k^2 \mathrm{Var}(x_k)$$

11. Mean and Variance of Sums
For any random variables,
$E(a_1 x_1 + \cdots + a_k x_k) = a_1 E(x_1) + \cdots + a_k E(x_k)$
For independent variables,
$\mathrm{Var}(a_1 x_1 + \cdots + a_k x_k) = a_1^2 \mathrm{Var}(x_1) + \cdots + a_k^2 \mathrm{Var}(x_k)$

12. Quantile

- x value at which CDF takes a value α is called α-quantile or 100α-percentile, denoted by x_α

$$P(x \leq x_\alpha) = F(x_\alpha) = \alpha$$

- If 90th-percentile score on GRE was 1500, then 90% of population got 1500 or less

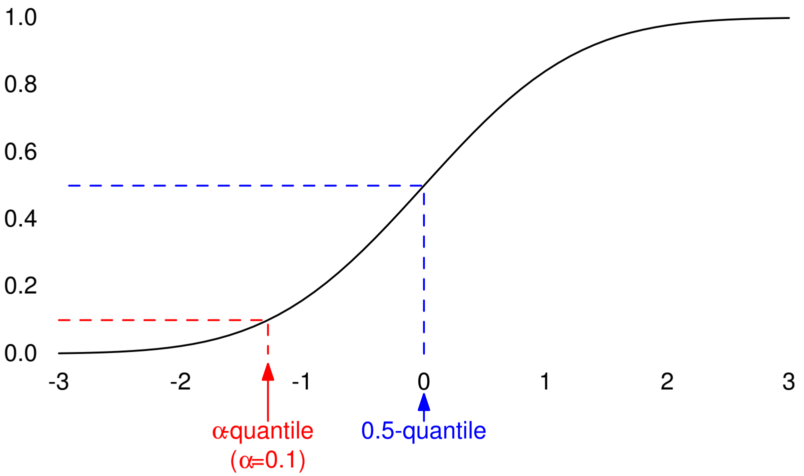
12. Quantile

- x value at which CDF takes a value α is called α-quantile or 100α-percentile, denoted by x_α

$$P(x \leq x_\alpha) = F(x_\alpha) = \alpha$$

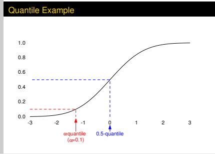
- If 90th-percentile score on GRE was 1500, then 90% of population got 1500 or less

Quantile Example



2015-06-15

CS147
└15 Concepts
 └Quantile
  └Quantile Example



13. Median

- 50th percentile (0.5-quantile) of a random variable
- Alternative to mean
- By definition, 50% of population is below median, 50% above
  - Lots of bad (good) drivers
  - Lots of smart (stupid) people

13. Median

- 50th percentile (0.5-quantile) of a random variable
- Alternative to mean
- By definition, 50% of population is below median, 50% above
  - Lots of bad (good) drivers
  - Lots of smart (stupid) people

14. Mode

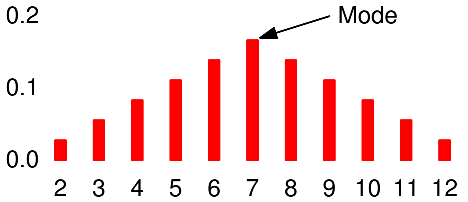
- Most likely value, i.e., x_i with highest probability p_i, or x at which pdf/pmf is maximum
- Not necessarily defined (e.g., tie)
- Some distributions are bi-modal (e.g., human height has one mode for males and one for females)

14. Mode

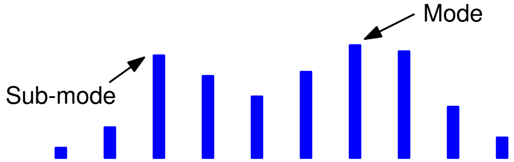
- Most likely value, i.e., x_i with highest probability p_i, or x at which pdf/pmf is maximum
- Not necessarily defined (e.g., tie)
- Some distributions are bi-modal (e.g., human height has one mode for males and one for females)

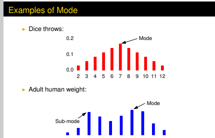
Examples of Mode

- Dice throws:



- Adult human weight:





15. Normal (Gaussian) Distribution

- Most common distribution in data analysis
- pdf is:

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

- $-\infty \leq x \leq +\infty$
- Mean is μ, standard deviation σ

15. Normal (Gaussian) Distribution
- Most common distribution in data analysis
- pdf is:
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$
- $-\infty \leq x \leq +\infty$
- Mean is μ, standard deviation σ

Notation for Gaussian Distributions

- Often denoted $N(\mu, \sigma)$
- Unit normal is $N(0, 1)$
- If x has $N(\mu, \sigma)$, $\frac{x-\mu}{\sigma}$ has $N(0, 1)$
- The α-quantile of unit normal $z \sim N(0, 1)$ is denoted z_α so that

$$\left\{P(\frac{x-\mu}{\sigma} \leq z_\alpha)\right\} = \{P(x) \leq \mu + z_\alpha \sigma\} = \alpha$$

Notation for Gaussian Distributions

- Often denoted $N(\mu, \sigma)$
- Unit normal is $N(0, 1)$
- If x has $N(\mu, \sigma)$, $\frac{x-\mu}{\sigma}$ has $N(0, 1)$
- The α-quantile of unit normal $z \sim N(0, 1)$ is denoted z_α so that

$$\left\{P(\frac{x-\mu}{\sigma} \leq z_\alpha)\right\} = \{P(x) \leq \mu + z_\alpha \sigma\} = \alpha$$

Why Is Gaussian So Popular?

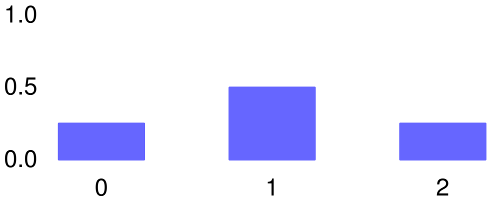
- We've seen that if $x_i \sim N(\mu_i, \alpha_i)$ and all x_i independent, then $\sum \alpha_i x_i$ is normal with mean $\sum \alpha_i \mu_i$ and variance $\sigma^2 = \sum \alpha_i^2 \sigma_i^2$
- Sum of large number of independent observations from any distribution is itself normal (Central Limit Theorem)
  $\Rightarrow$ Experimental errors can be modeled as normal distribution.

Why Is Gaussian So Popular?

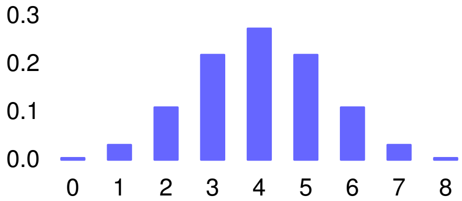
- We've seen that if $x_i \sim N(\mu_i, \alpha_i)$ and all x_i independent, then $\sum \alpha_i x_i$ is normal with mean $\sum \alpha_i \mu_i$ and variance $\sigma^2 = \sum \alpha_i^2 \sigma_i^2$
- Sum of large number of independent observations from any distribution is itself normal (Central Limit Theorem)
  $\Rightarrow$ Experimental errors can be modeled as normal distribution.

Central Limit Theorem

- Sum of 2 coin flips (H=1, T=0):



- Sum of 8 coin flips:



2015-06-15

CS147
└15 Concepts
  └Normal Distribution
    └Central Limit Theorem