NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W3: 9/1-3/15

NC STATE UNIVERSITY

Administrative

- Waiting to get into class – today is the last day
- 1st H/W
  - Posted by 9/7/15
  - Due (2 weeks): 9/21/15
- Books
  - https://www.openintro.org/stat/

NC STATE UNIVERSITY

Key points from 8/27

- Experiment, outcomes, sample space, events
- Basic set operations
- Probability and three axioms
- Probability rules
- Independent events
- Conditional probability
- Bayes theorem
- Random variables: discrete and continuous
- PMF and Bernoulli distribution

NC STATE UNIVERSITY

Today

- Continuous Probability Distributions

NC STATE UNIVERSITY

Probability Density Function

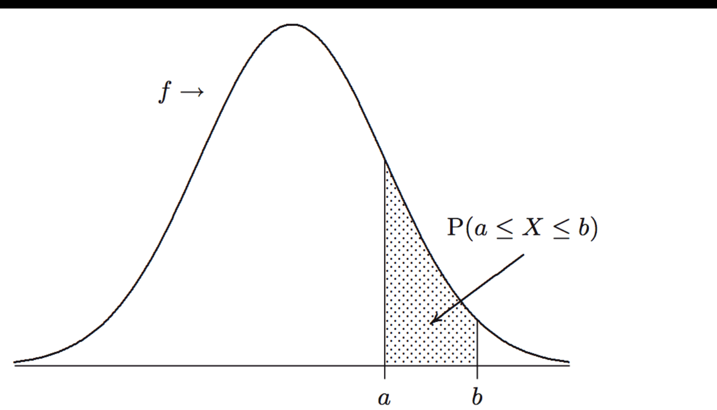
DEFINITION. A random variable X is *continuous* if for some function $f : \mathbb{R} \to \mathbb{R}$ and for any numbers a and b with $a \leq b$,

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x.$$

The function f has to satisfy $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. We call f the *probability density function* (or *probability density*) of X.

NC STATE UNIVERSITY

Area



Area under a probability density function f on the interval $[a, b]$.

Probability that X lies in an interval [a, b] is equal to the area under the probability density function f of X over the interval [a,b]

What if the interval is small

- If the interval get small and small, then what will be the probability?

$$P(a - \varepsilon \leq X \leq a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,dx$$

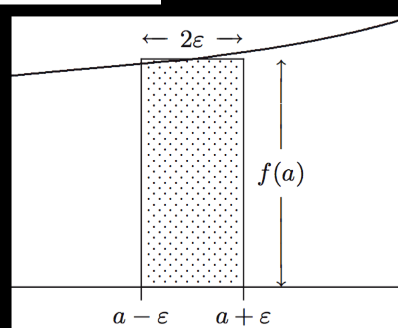
- If ε-> 0, then for any "a", P(X=a) = 0

9/2/15 © Raju Vatsavai CSC-591. 7

What does f(a) means?

$$P(a - \varepsilon \leq X \leq a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,dx \approx 2\varepsilon f(a)$$

- f(a) can be interpreted as a (relative) measure of how likely it is that X will be near a



Approximating the probability that X lies ε close to a

9/2/15 © Raju Vatsavai CSC-591. 8

NC STATE UNIVERSITY

Exercise

- Let f(x) = 0 if x ≤ 0 or x ≥ 1, and f(x) = 1/(2 x sqrt(x)) for 0 < x < 1. First verify that it satisfies two properties of pdf. Let X be a random variable with f as its pdf. Compute the probability that X lies between 10^{-4} and 10^{-2}

NC STATE UNIVERSITY

Solution

- We know from integral calculus that 0 ≤ a ≤ b ≤ 1, we have

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b \frac{1}{2\sqrt{x}}\,\mathrm{d}x = \sqrt{b} - \sqrt{a}.$$

NC STATE UNIVERSITY

Continuous Uniform Distribution

DEFINITION. A continuous random variable has a *uniform distribution* on the interval $[\alpha, \beta]$ if its probability density function f is given by $f(x) = 0$ if x is not in $[\alpha, \beta]$ and

$$f(x) = \frac{1}{\beta - \alpha} \quad \text{for } \alpha \leq x \leq \beta.$$

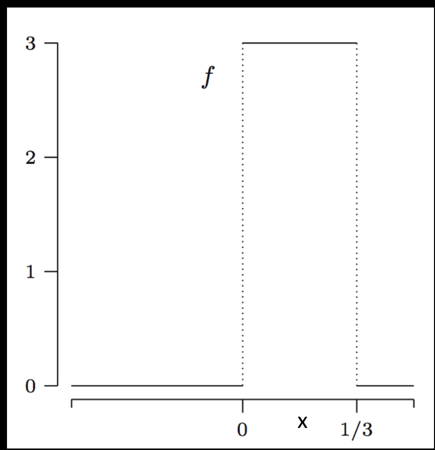
We denote this distribution by $U(\alpha, \beta)$.

NC STATE UNIVERSITY

pdf of U(0, 1/3)

U(a,b) => is constant over possible values of x

Any equal length intervals in (ab, b) are also equally likely



NC STATE UNIVERSITY

Example

A = B X H = (d-c)f(x) = 1
H = f(x) = 1/(d-c)

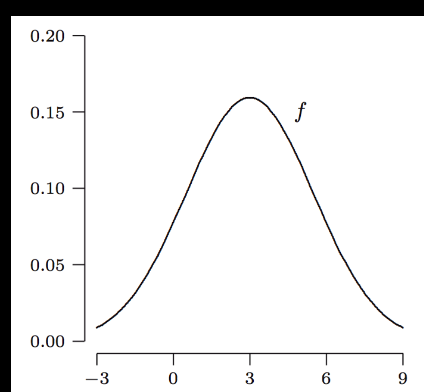
Say c = 100, d = 150
f(x) =?
P(X > 125)
25 percentile

NC STATE UNIVERSITY

Normal Distribution

- Also known as Gaussian distribution
- Is an extremely important distribution

N(3,6.25)



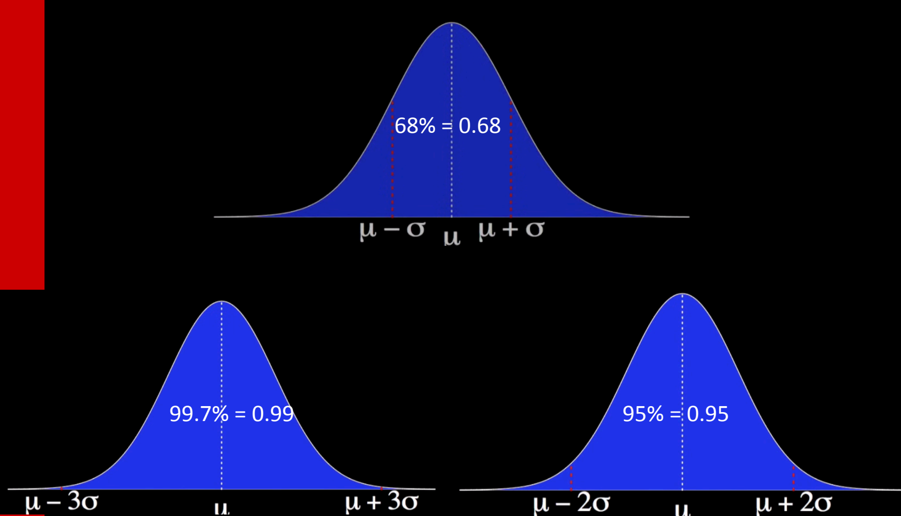
Normal Distribution

DEFINITION. A continuous random variable has a *normal distribution* with parameters μ and $\sigma^2 > 0$ if its probability density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty.$$

We denote this distribution by $N(\mu, \sigma^2)$.

Key points



NC STATE UNIVERSITY

Quantile

DEFINITION. Let X be a continuous random variable and let p be a number between 0 and 1. The pth *quantile* or $100p$th *percentile* of the distribution of X is the smallest number q_p such that

$$F(q_p) = \mathrm{P}(X \leq q_p) = p.$$

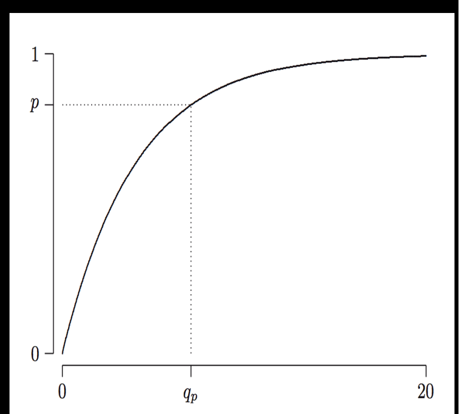
The *median* of a distribution is its 50th percentile.

9/2/15 © Raju Vatsavai CSC-591. 17

NC STATE UNIVERSITY

q_p

- For crv, q_p is easy to compute. F is strictly increasing from 0 to 1 on some interval.
- $q_p = F^{-1}(p)$
- For exponential distributions, its easy to compute, but for Normal distribution, we need a lookup table



The p^{th} q_p of exp(0.25) distribution

9/2/15 © Raju Vatsavai CSC-591. 18

NC STATE UNIVERSITY

Example

- Find 0.95th quantile of standard normal distribution

NC STATE UNIVERSITY

Acknowledgements

- Vipin Kumar (Minnesota)
- Jiawei Han (UIUC)
- Hanspeter Pfister (Harvard)
- Larry Wasserman (CMU)