Statistical Distributions and Uncertainty Analysis

QMRA Institute

Patrick Gurian

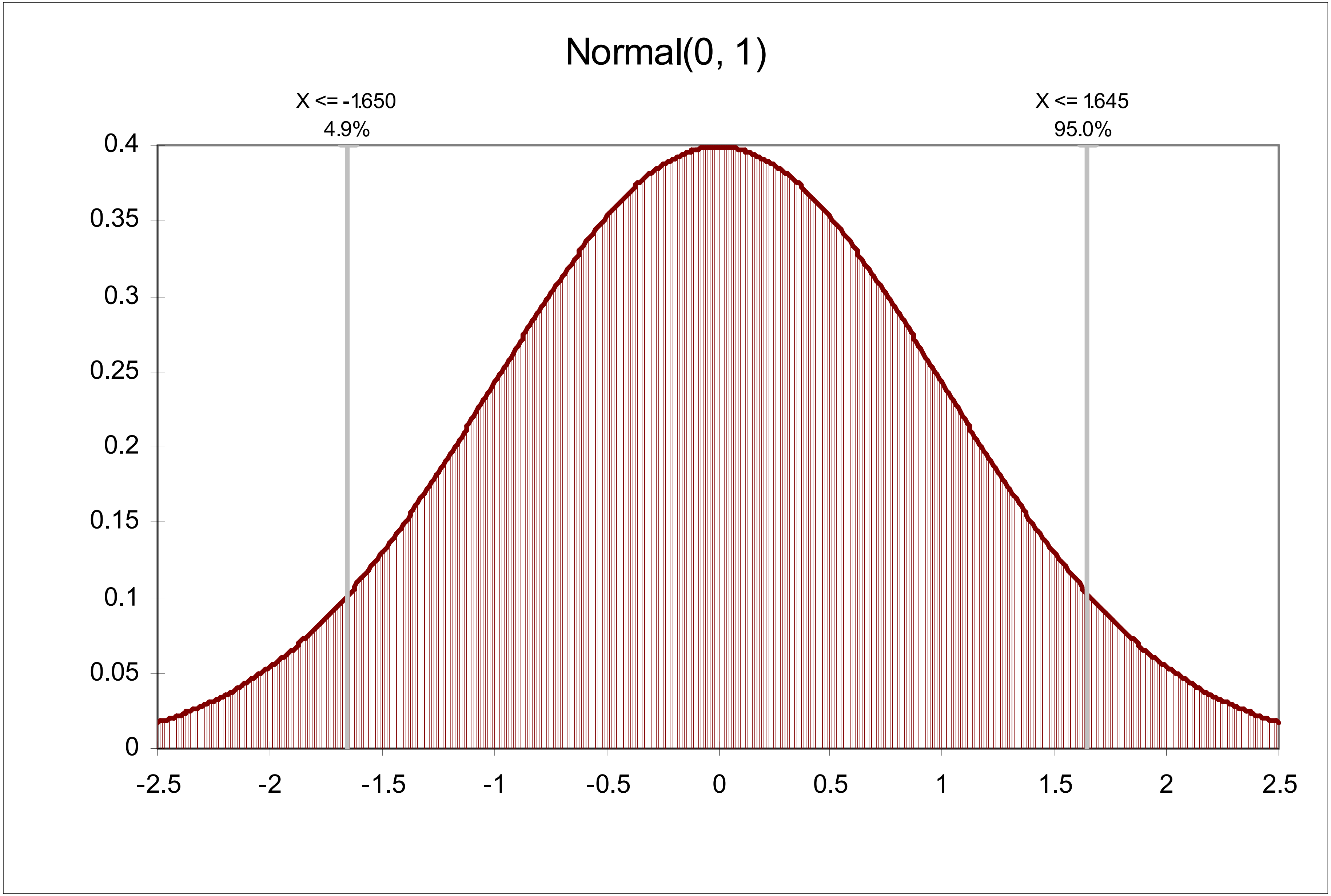
Probability

- Define a function f(x) probability density distribution function (PDF)
- Prob [A<x<B] = $\int_A^B f(x)dx$

Parameters

- f(x) usually has a number of constants
- These constants can be "tuned" to fit different applications
- Model parameters
- For normal distribution
- $f(x) = 1/\{\sigma(2\pi)^{1/2}\} \exp\{(x-\mu)^2/ 2\sigma^2\}$
- Pick any μ and σ to define a normal PDF
- Notation is $x \sim N(\mu, \sigma^2)$

Standard Normal Distribution



CDF

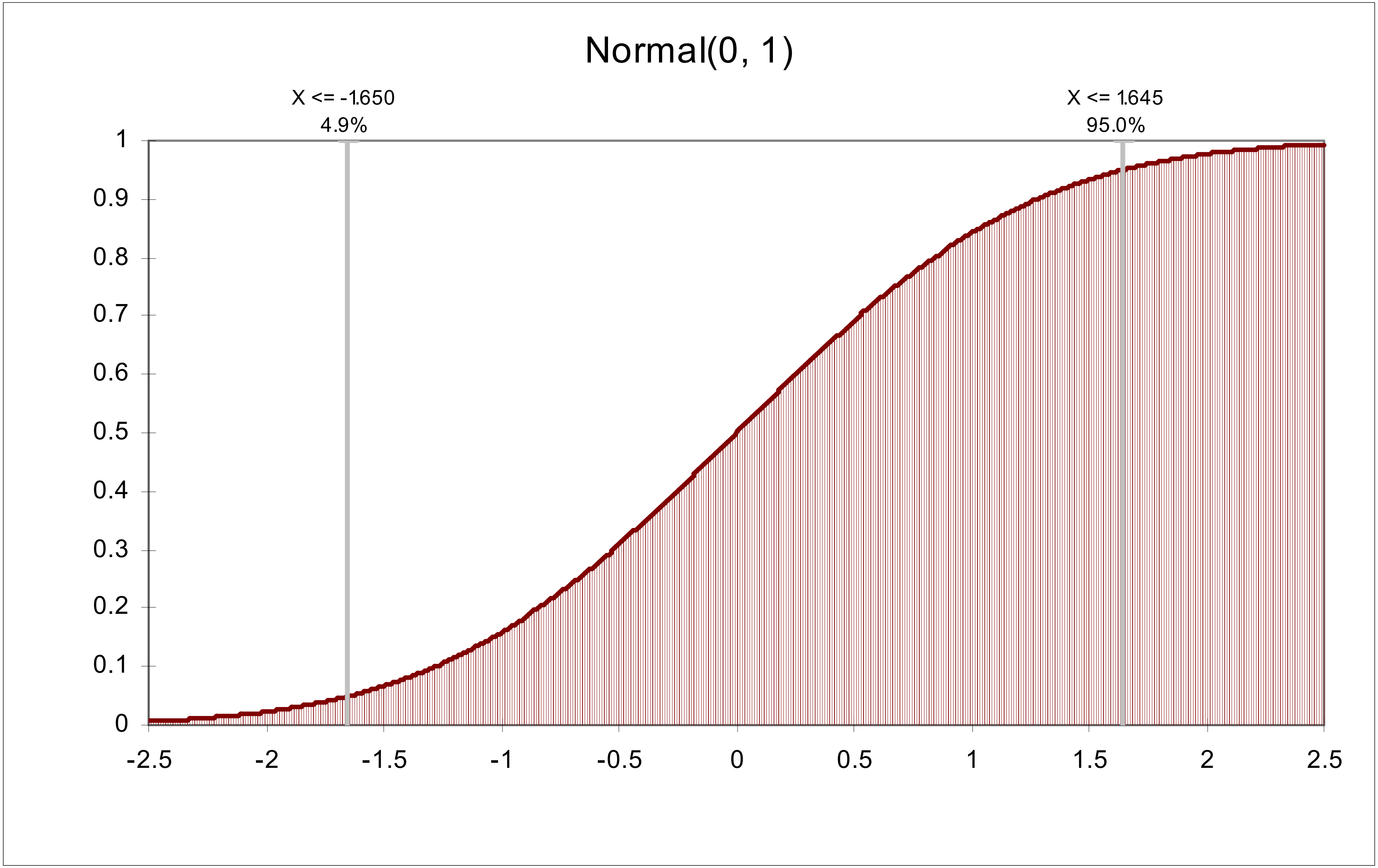
- $\int^{X} f(x)dx = F(X)$
- F(x) is the cumulative distribution function (CDF)

- Prob[A<x<B] = F(B) - F(A)

- $\int^{m} f(x)dx = F(m)=0.5$
- What is m?

- For a normal distribution $\int^{x} f(x)dx$ requires numerical integration

Normal CDF



Normal(0, 1)
X <= -1.650
4.9%
X <= 1.645
95.0%
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

Short cut for normal F(X)

- Define $Z = (X-\mu)/\sigma$
- $Z \sim N(0, 1)$
- For $\sigma>0$ this is a monotonic transformation
- The largest X corresponds to the largest Z
- The 95th percentile of Z corresponds to the 95th percentile of X
- Tabulate Z

Example

- Car repairs are ~ N $(300, 100^2)$
- What is prob [repair > 450]?
- $Z = (X-\mu)/\sigma$
- $Z = (450-300)/100 = 1.5$
- $F(Z) = F(1.5) = 0.933$
- see table of Z values
- $0.933 = \text{prob}[Z<1.5] = \text{prob}[\text{repair}<450]$
- Prob [repair >450]=0.067

Normal distribution in Excel

=Normdist (x, mean, st. dev., cumulative)

=Norminv (probability, mean, st. dev.)

- Norminv is always cumulative by definition

Variability

- Changes in outcome over time, space, or different trials
- Concentration of microbes in water samples
- Amount of water consumed
  - Even if we know everything about the system there is variability in how much water different people drink

Uncertainty

- Uncertainty is a lack of knowledge
- Risk presented by low doses of toxins
- Sensitivity of the climate system to a doubling of pre-industrial carbon dioxide
  - There is only one value, we just don’t know it

Variable and Uncertain

- A limited amount of data is available on the concentration of Cryptosporidium in drinking water
- Based on N=20, mean = 3/100 liter, variance =3/100 liter
- The number in any given sample is variable with a Poisson distribution describing this variability
- The true mean is uncertain
  – we just estimated it based on a limited sample

Probability Distributions as Models

- Probability was originally developed for describing variability
- It is now widely applied to describe uncertainty

Bayesian Framework

- Describe variability in observations with a probability distribution
- Stage 1: oocysts ~ Poisson(λ)
- Describe uncertainty in this parameter a probability distribution
- Stage 2: $\ln(\lambda) \sim N(\mu, \sigma)$

Fitting Distributions

- We often need to find probability distributions to describe variability and uncertainty in risk assessments
- Often start with a general class of model and then adjust it to match the specific case

Statistical Model Estimation

- Statistical models all contain parameters
- Parameters are constants that can be “tuned” to make a general class of models applicable to a specific dataset

Example: Normal distribution

PDF is:

$f(x) = 1/\{\sigma(2\pi)^{1/2}\} \exp\{(x-\mu)^2/ 2\sigma^2\}$

μ and σ are constants that define a particular normal distribution

We want to pick values that match a given dataset

One simple way is arithmetic mean=μ

s=σ (method of moments)

But there are other ways including…

Maximum likelihood estimation

- Assume a probability model
- Calculate the probability (likelihood) or obtaining the observed data
- Now adjust parameter values until you find the values that maximize the probability of obtaining the observed data

Likelihood Function

- Observe data x, a vector of values
- Assume some pdf $f(x|\theta)$
- Where θ is a vector of model parameters
- Probability of any particular value is $f(x_i| \theta)$ where i is an index indicating a particular observation in the data

Likelihood of the data

- Generally assume data is independent and identically distributed
- Same f(x) for all data
- For independent data:

prob [Event A ∩ Event B] = prob[A] Prob[B]

- So multiple probability of individual observations to get joint probability of data
- $L=\pi\ f(x_i|\ \theta)$
- Now find θ that maximizes L

MLE example

- A team plays 3 games: W L L
- Binomial model: what is p?
- L=(3:1)p(1-p)^2

- Suppose we know sequence of wins and losses then we can say
- L=p(1-p)^2

MLE example (cont)

- Suppose p=.5
- L=0.5^2*0.5=0.125

Example for class

- Now suppose p=.3
- What is likelihood of data?
- Do we prefer p=0.5 or p=0.3?

Answer

Example: Maximizing the likelihood

$dL/dp = 3(1-p)^2 - 6p(1-p)$

Find maximum at $dL/dp=0$

$0=3(1-p)^2 - 6p(1-p)$

$0=1-p -2p$

$3p=1$

$p=1/3$

MLE example: Conclusion

Can verify that this is a maxima by looking at second derivative

Note that method of moments would give us $p=x/n=1/3$

So we get the same result by both methods

Ln Likelihood

- Product of many numbers each <1 is quite small
- Often easiest to work with ln (L)
- Since ln is a monotonic transformation of L, the largest ln (L) will correspond to the largest L value

$\text{Ln } L = \ln \pi\, f(x_i | \theta)$

Applying log laws

$\text{Ln } L = \Sigma \ln f(x_i | \theta)$

Uncertainty in parameter estimates

- Generally statistical methods quantify sample variability ***AND ONLY*** sample variability
- The properties of the specific sample that I took will vary from the long run population properties
- But as my sample becomes large its properties approach those of the population
- Statistical methods quantify how much I know about the population given a particular sample

Standard Errors

- Standard error is the variance of the parameter estimate
- For simple cases formulae exist for standard errors of parameter estimates
- Arithmetic mean ~ $N(\mu, \sigma^2/n)$
  - Standard error is σ^2/n
- Formulae are not always available

Bootstrapping

- A general approach to generating confidence intervals for any statistic
- Assume your sample is a discrete distribution, PMF for population
- Sample from this PMF with replacement and get observed distribution of statistics of interest
- Generate a new sample set of same size as original sample size since you usually want confidence interval for this size of sample
- Calculate statistic of interest
- Repeat (many times) and characterize distribution of statistic of interest

Basis of bootstrap

- Because you sample with replacement you get a randomly varying different sample
- Makes sense intuitively
- Developed and used and then later justified by theory

Example for class

- What is the probability that in n samples from a sample of size n, a given data point is not sampled?
- Prob that this data point is sampled each time is?
- 1/n

Solution

Bootstrap approach for upper bound on exposure

- Bootstrap a number of samples
- Find the upper bound you are interested in for each sample
- Now find the distribution of upper bounds, select a probability interval for the upper bound

Bootstrap advantages

- Bootstrap will give you not just upper bound but also uncertainty range for this upper bound
- Generally applicable
  - No distributional assumptions