STEVENS INSTITUTE of TECHNOLOGY
School of Business

BIA-652

Multivariate Data Analytics

Random Variables (continued)

Prof. Feng Mai
School of Business

For academic use only.

Continuous Random Variable

Continuous Random Variable

- Model of probabilities attached to events.
- Probability Density Function (PDF), f(x)
- Probability of an event P(event) obtained by integrate the PDF over a range.

Density: $f(x) \geq 0$

Definition: $\int_{-\infty}^{\infty} f(x)dx=1$

Definition: $CDF=Prob(X \leq x)=F(x)=\int_{-\infty}^{x} f(x)dx$

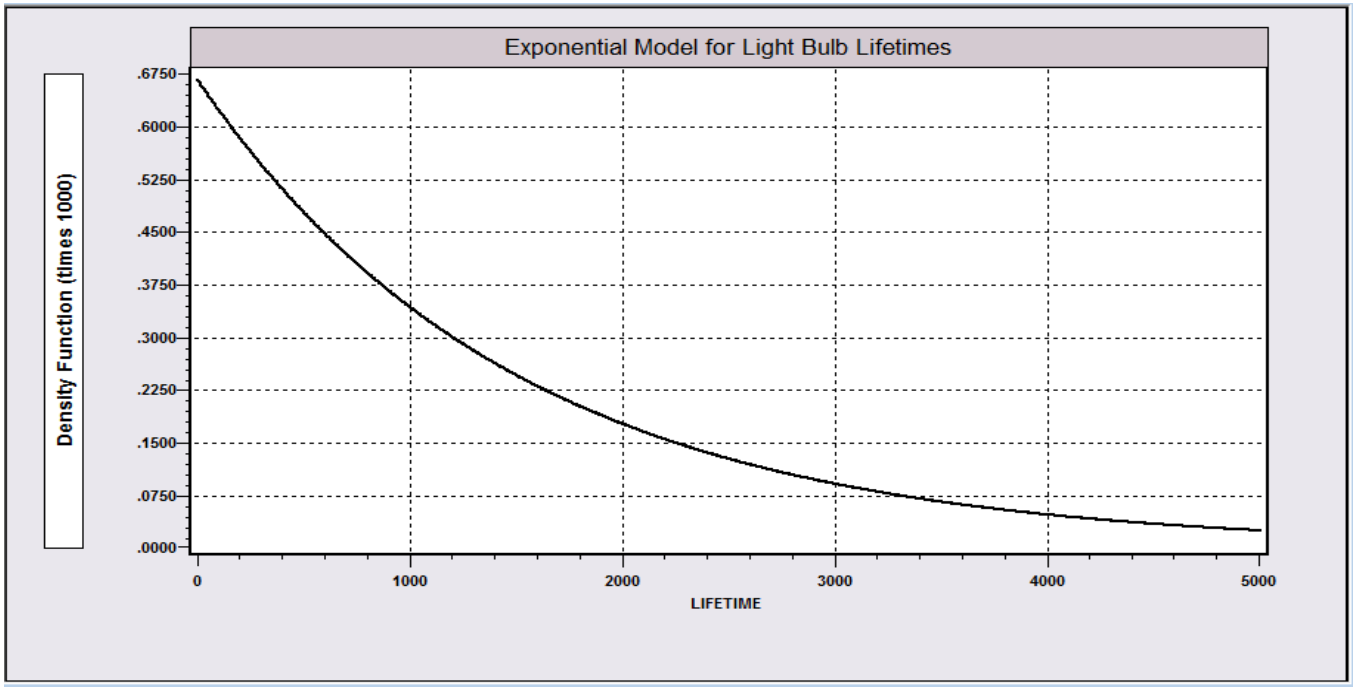
('Cumulative Density Function' or 'Distribution Function')

Probability: $P(\omega)=\int_{\omega} f(x)dx$

In range a to b $Prob(a < X < b) = Prob(a \leq X \leq b) = \int_{a}^{b} f(x)dx$

An Exponential Model for Light Bulb Lifetimes

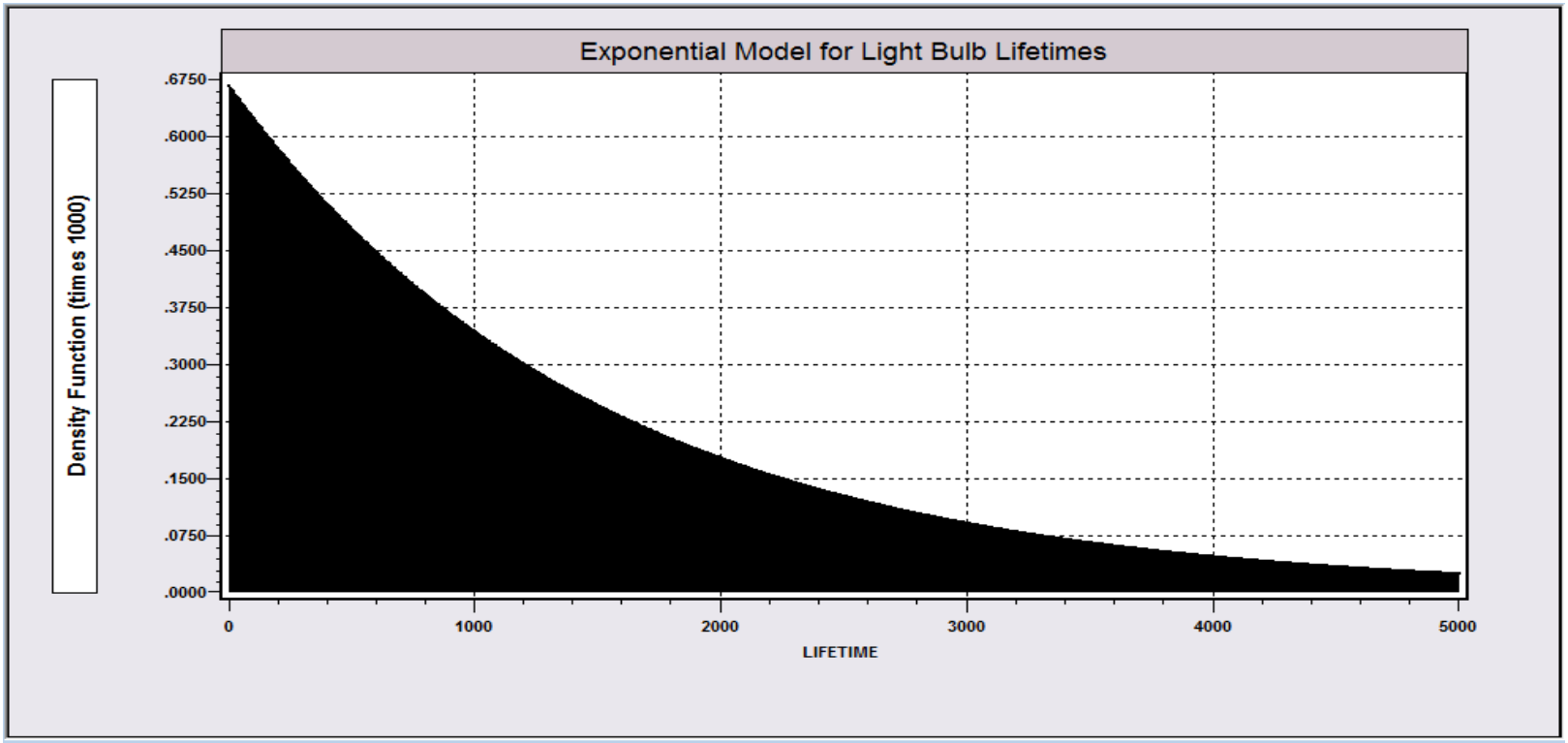
- This is the exponential model for lifetimes. The PDF is $f(x) = (1/\mu)\, e^{-x/\mu}$
- The model parameter $\mu = 1500$
- Alternatively, it can be written as $f(x) = \lambda e^{-\lambda x}$, $\lambda = 1/1500$



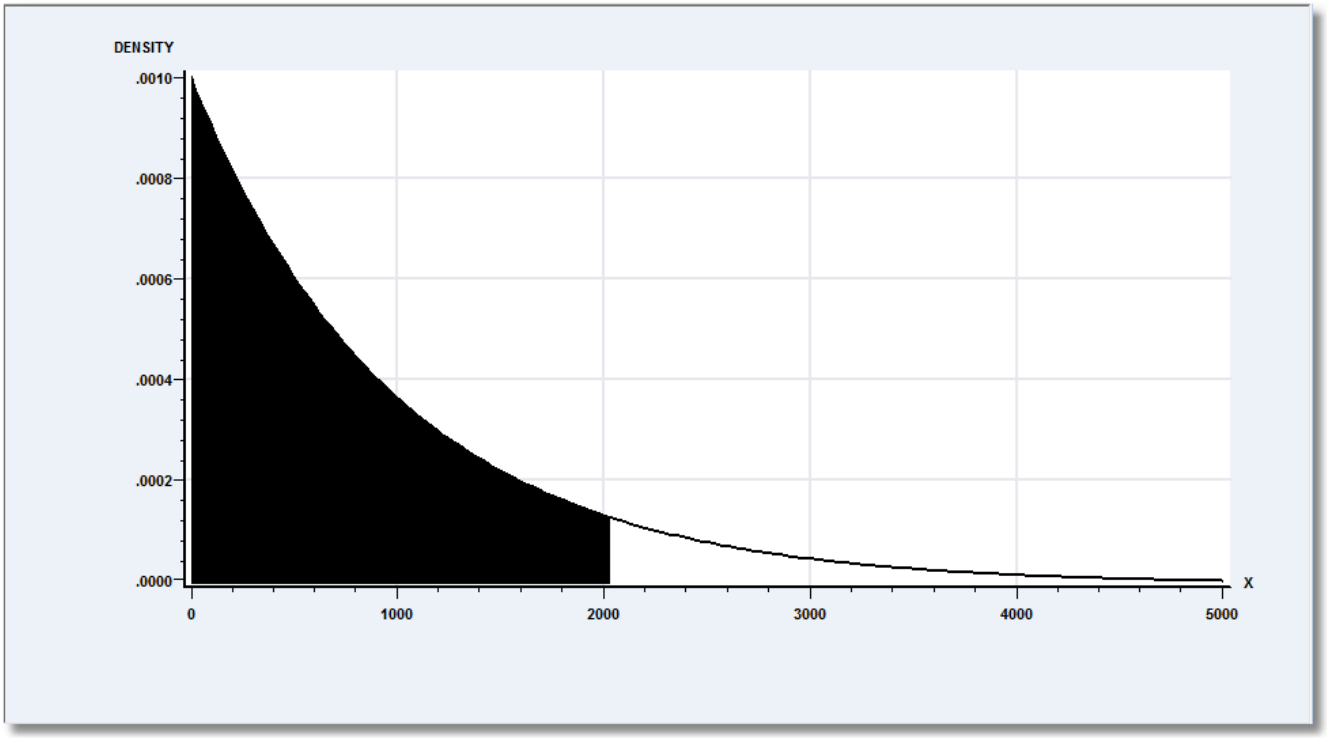
BIA 652, Multivariate Data Analytics

Model for Light Bulb Lifetimes

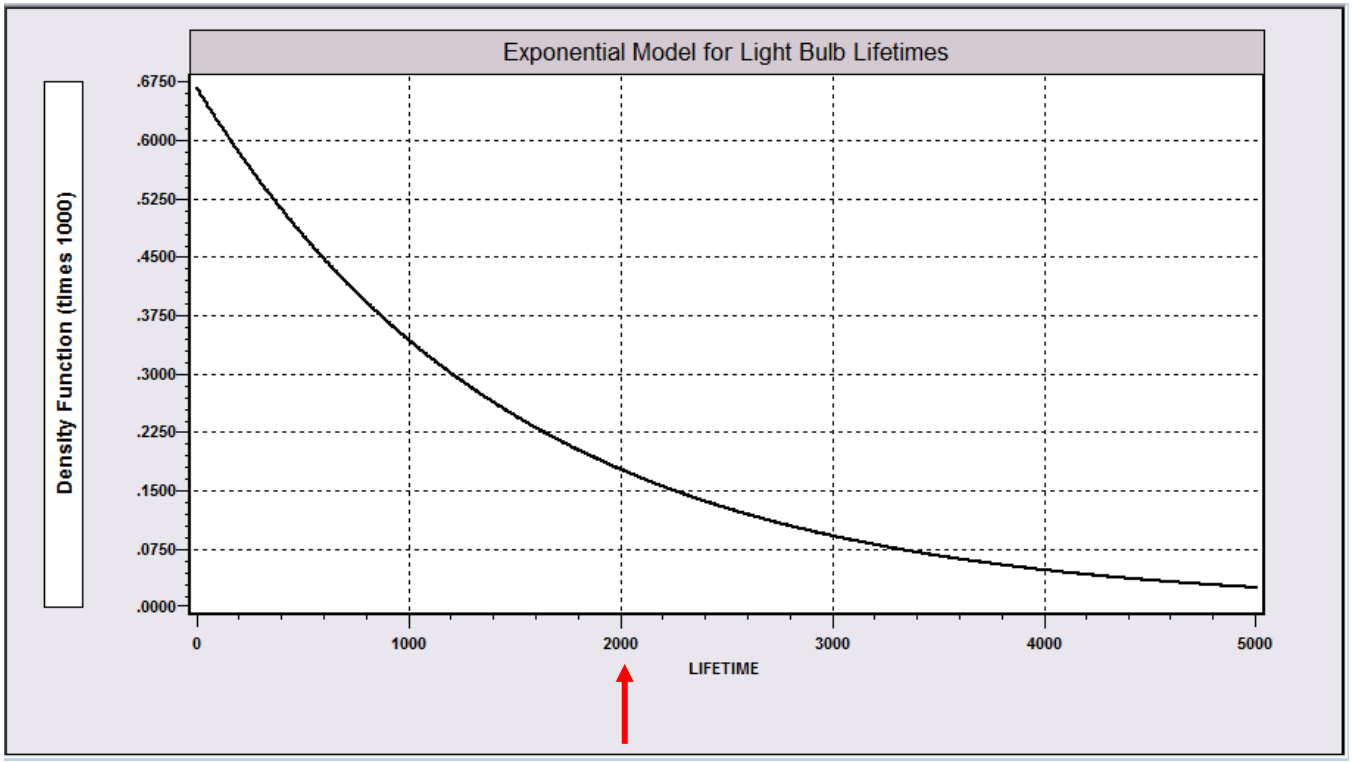
The area under the entire curve is 1.0.



The CDF is Prob(X $\leq$ x) = the area under the density to the left of x. This is P(X $\leq$ 2000).



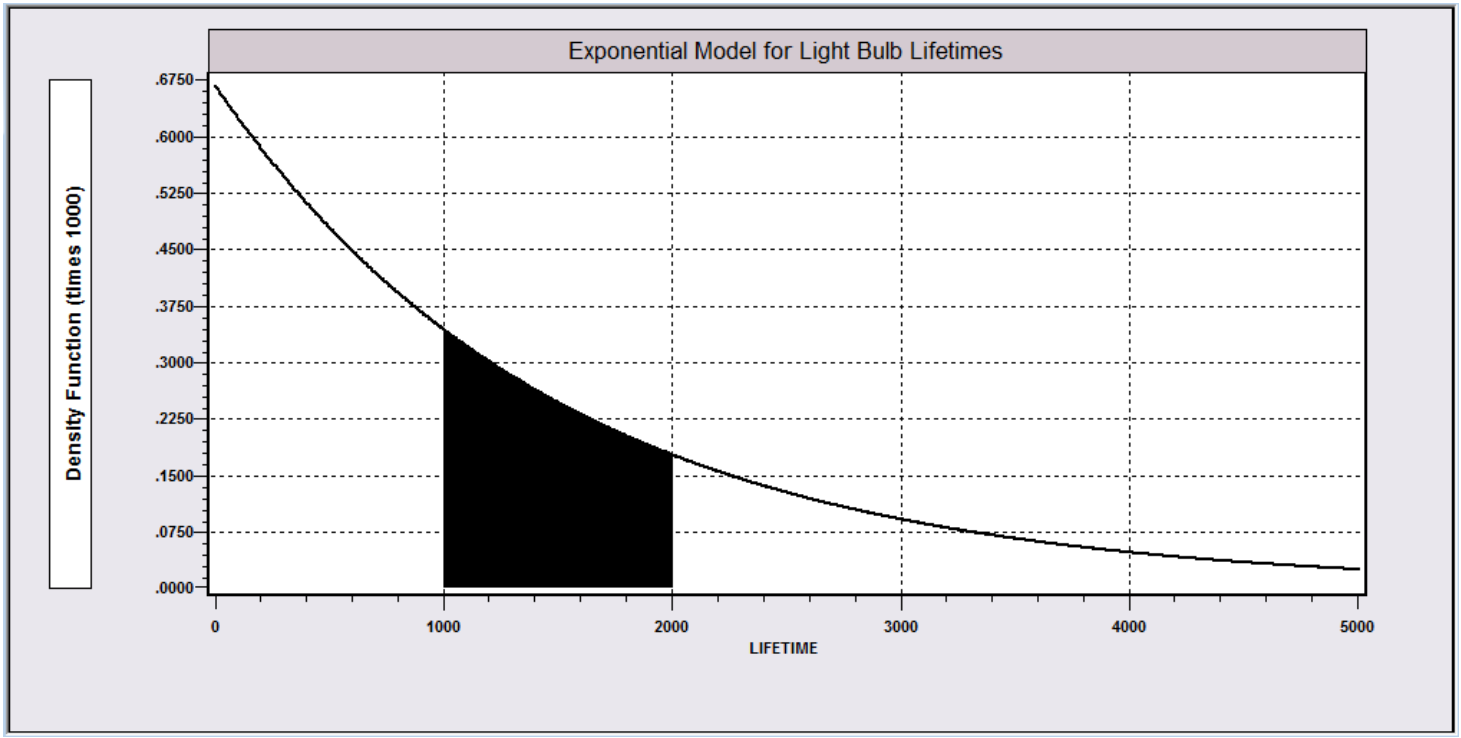
Probability of a Single Value Is Zero



The probability associated with a single point, such as LIFETIME=2000, equals 0.0.

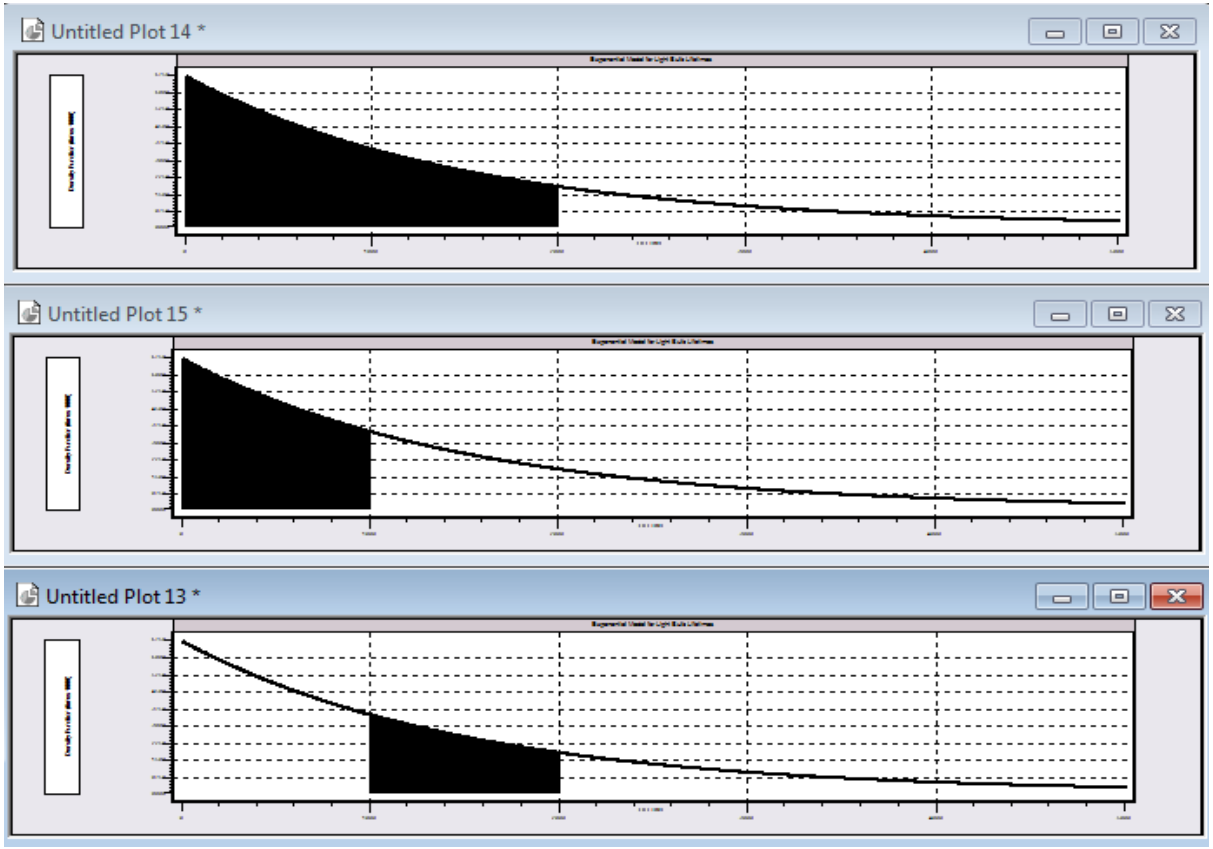
Continuous Distribution

The probability associated with an interval such as 1000 < LIFETIME < 2000 equals the area under the curve from the lower limit to the upper.



Probability for a Range of Values Based on CDF

$$\text{Prob}(a \le X \le b) = \int_a^b f(x)dx$$
$$= F(b) - F(a)$$



- Prob(Life ≤ 2000) (.7364)

Minus

- Prob(Life ≤ 1000) (.4866)

Equals

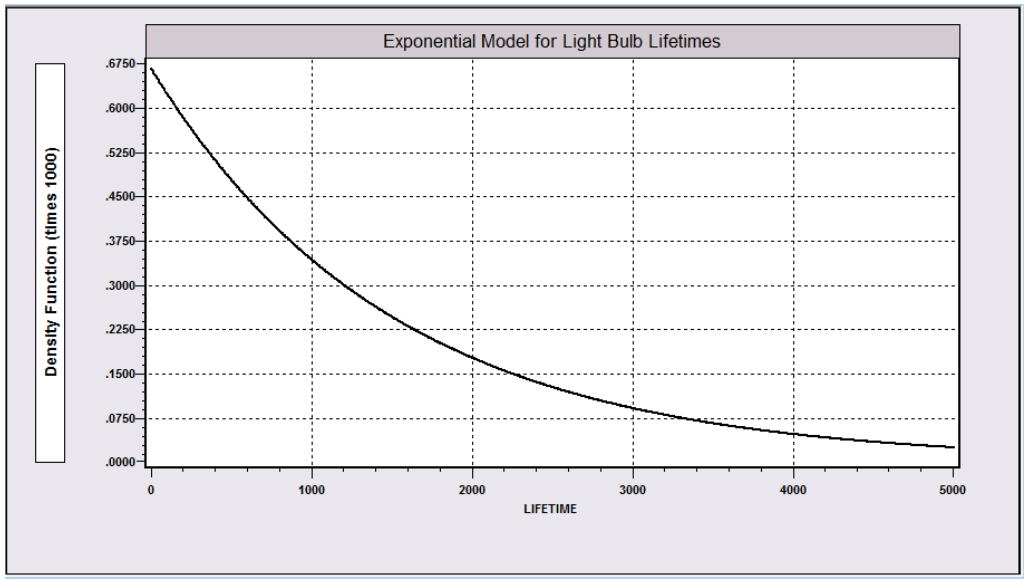
- Prob(1000 ≤ Life ≤ 2000) = (.2498)

CDF and Quantiles

- pth = quantile; $0 < p < 1$
- Pth Quantile = x_p such that $F(x_p) = p$.
- $x_p = F^{-1}(p)$.
- For $p = .5$, x_p = median

Median of Exponential Distribution

- $f(x) = \lambda \exp(-\lambda x)$, $x \geq 0$, 0 otherwise
- $F(x) = 1 - \exp(-\lambda x)$, $x \geq 0$



Median: $F(M)$ $= .5$

$1 - \exp(-\lambda M) = .5$

$\exp(-\lambda M) = .5$

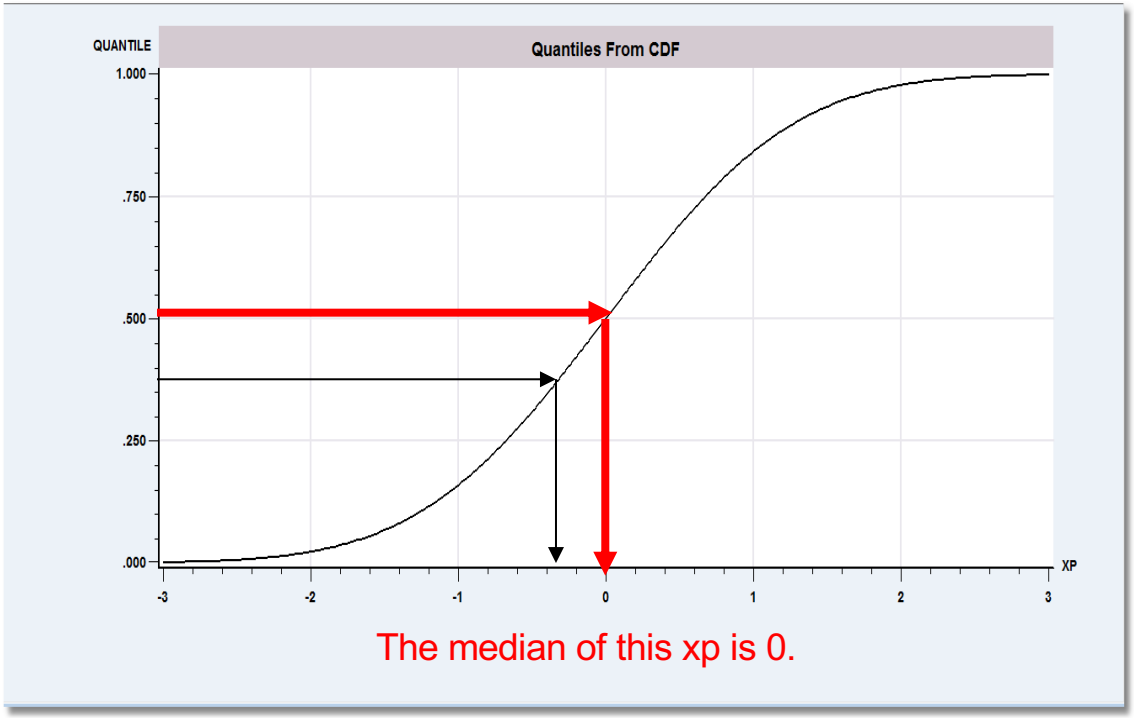
$-\lambda M = \ln.5$

$M = -\ln.5/\lambda$

$= (\ln 2)/\lambda$

Quantile = Prob(X ≤ xp)

This is p = F(xp)



This is xp = $F^{-1}(p)$

A **uniform random variable** on the interval [0, 1] is a model for what we mean when we say, "choose a number at random between 0 and 1."

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & x < 0 \text{ or } x > 1 \end{cases}$$

CDF:

X ~ Uniform[0, 1]

probability that x < 0.3?

median of x?

Common Continuous RVs

- Continuous random variables are all models; they do not occur in nature. The usual model builder's toolkit:
  - Continuous uniform
  - Exponential
  - Normal
  - Lognormal
  - Gamma
  - Beta
- Defined for specific types of outcomes
- There are thousands of other models

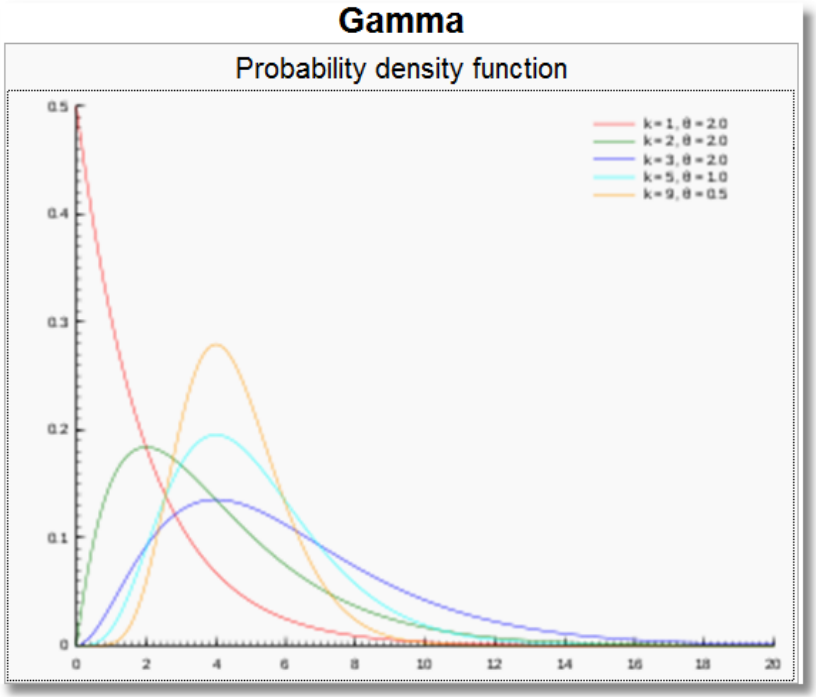
Representations of a Continuous Random Variable

- Representations
  - Density, f(x)
  - CDF, F(x) = Prob(X $\leq$ x)
  - Survival, S(x) = Prob(X $\geq$ x) = 1-F(x)
  - Hazard function, h(x) = -dlnS(x)/dx

- Representations are one to one – each uniquely determines the distribution of the random variable

Gamma Distributed Random Variable

$$f(x \mid \lambda, P) = \frac{\lambda^P x^{P-1} e^{-\lambda x}}{\Gamma(P)}, x \geq 0,\ \lambda > 0,\ \mathrm{P} > 0.$$



- Used to model nonnegative random variables – e.g., survival of people and electronic components
- Special cases
- P = 1 is the exponential distribution
- P = ½ and λ = ½ is the chi squared with one “degree of freedom”
- P = n/2 and λ = ½ is the chi squared distribution with n degrees of freedom

Beta Uses Beta Integrals

Beta density is used to model a random variable that ranges from 0 to 1 such as a proportion

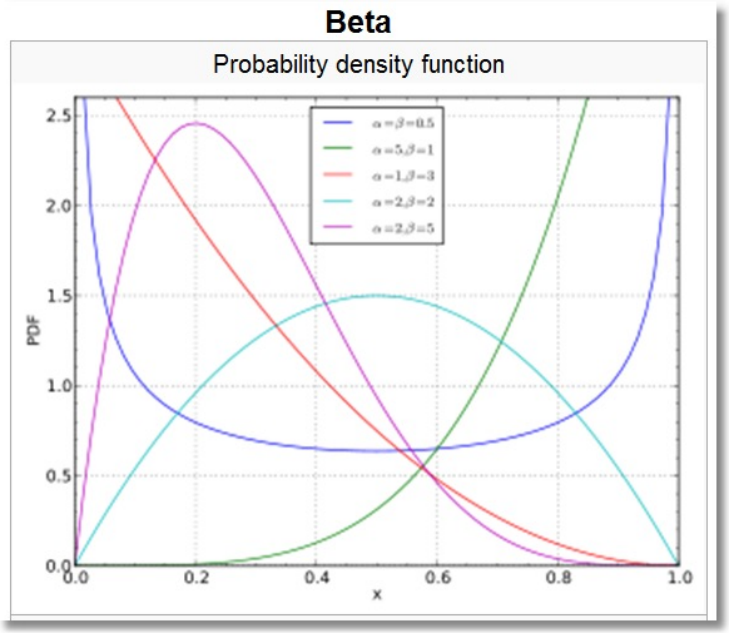
$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

The beta density is $f(x|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$,

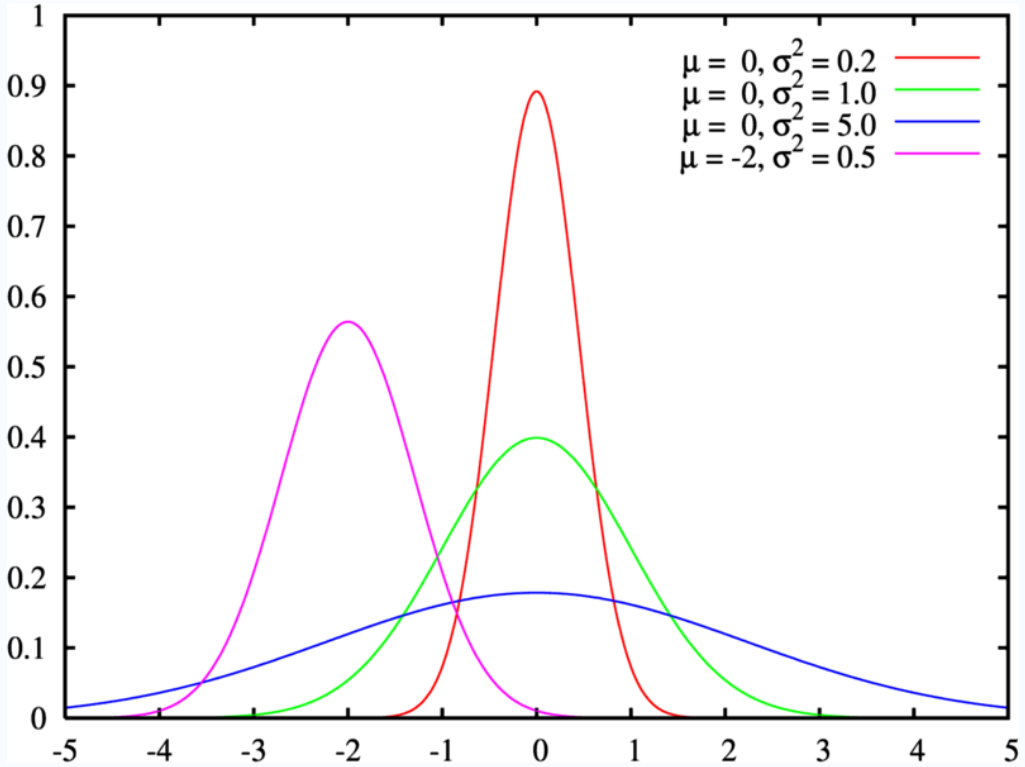
$0 \le x \le 1, a > 0, b > 0.$

The shape of the density depends on a and b.

Useful special case, a = 1 and b = 1 is the Uniform(0,1)



Normal Distributions



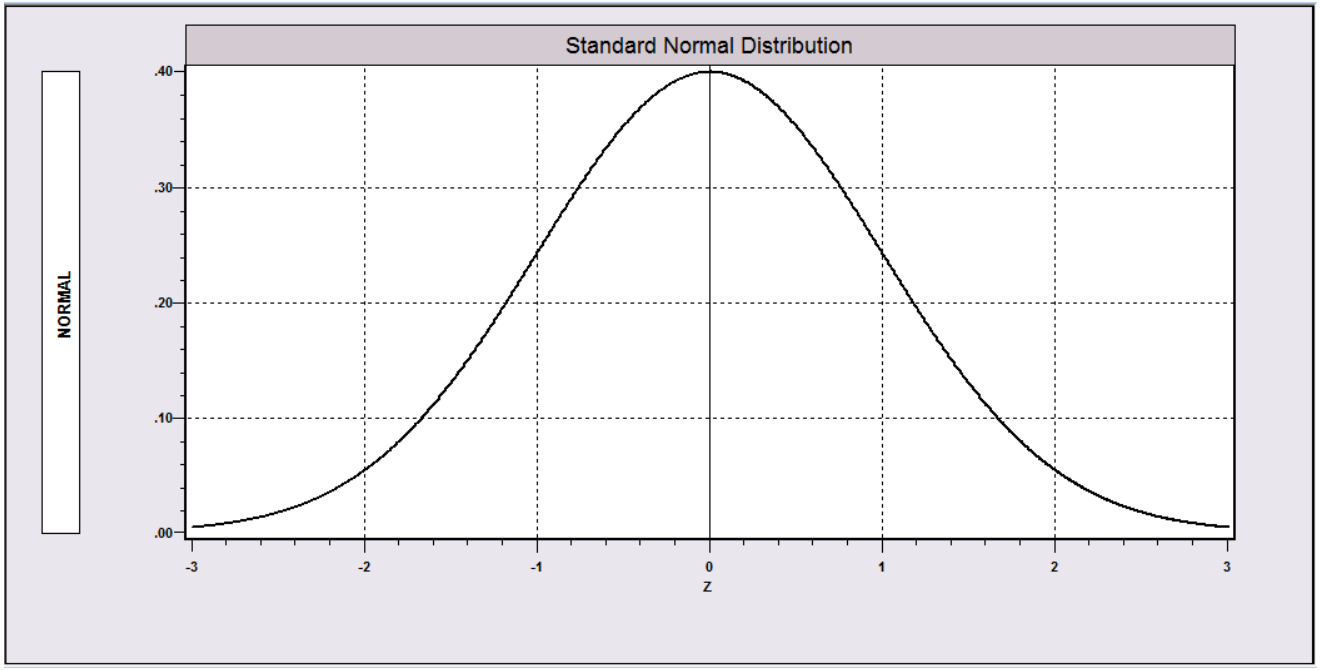
The scale and location (on the horizontal axis) depend on μ and σ. The shape of the distribution is always the same. (Bell curve)

Normal (Gaussian) Density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right], -\infty < x < +\infty$$

Mean = μ, standard deviation = σ

Standard Normal Density (μ=0, σ=1)



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

CDF of Standard Normal

Z ~ N(0,1) "*standard* (or *unit*) normal"

Use $\Phi(z)$ to denote CDF, i.e.

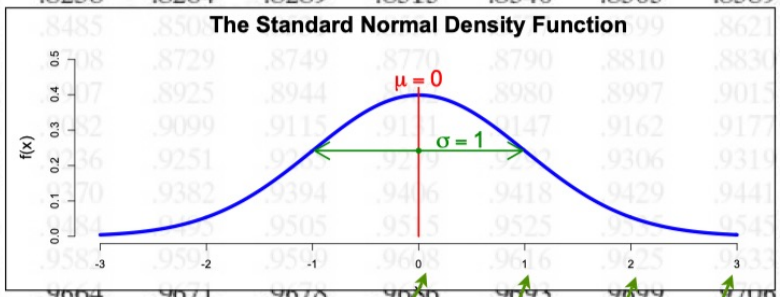
$$\Phi(z) = \Pr(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

no closed form ☹

TABLE 5.1: AREA $\Phi(x)$ UNDER THE STANDARD NORMAL CURVE TO THE LEFT OF X

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$\Phi(.46)$



BIA 652, Multivariate Data Analytics

Standard Normal Distribution Facts

- The random variable z runs from $-\infty$ to $+\infty$
- $\phi(z) > 0$ for all z, but for $|z| > 4$, it is essentially 0.
- The total area under the curve equals 1.0.
- The curve is symmetric around 0. (The normal distribution generally is symmetric around μ.)

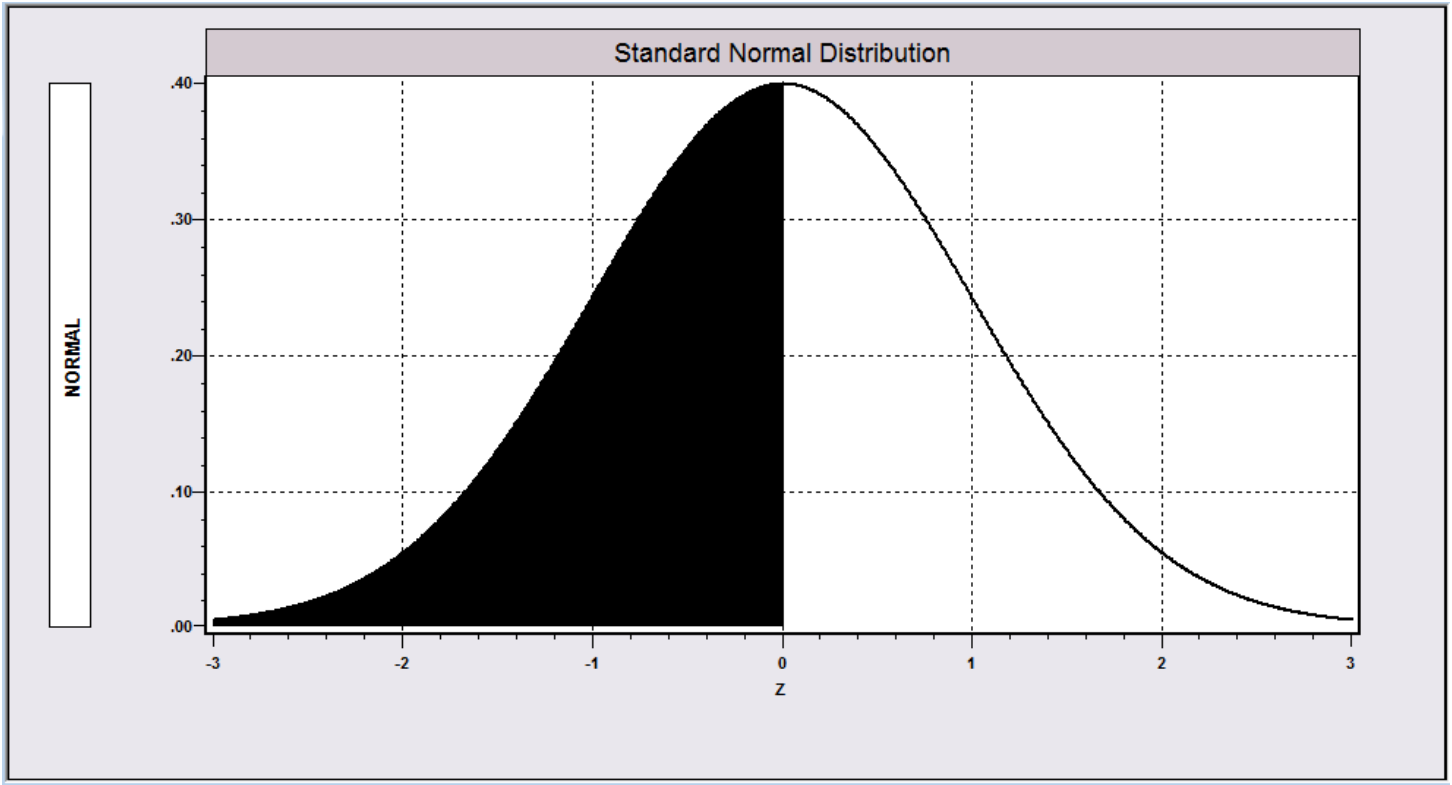
Computing Probabilities

- Standard Normal Tables give probabilities when $\mu = 0$ and $\sigma = 1$
- For other cases, do we need another table?
- Probabilities for other cases are obtained by “standardizing.”
  - Standardized variable is $z = (x - \mu)/\sigma$
  - z has mean 0 and standard deviation 1

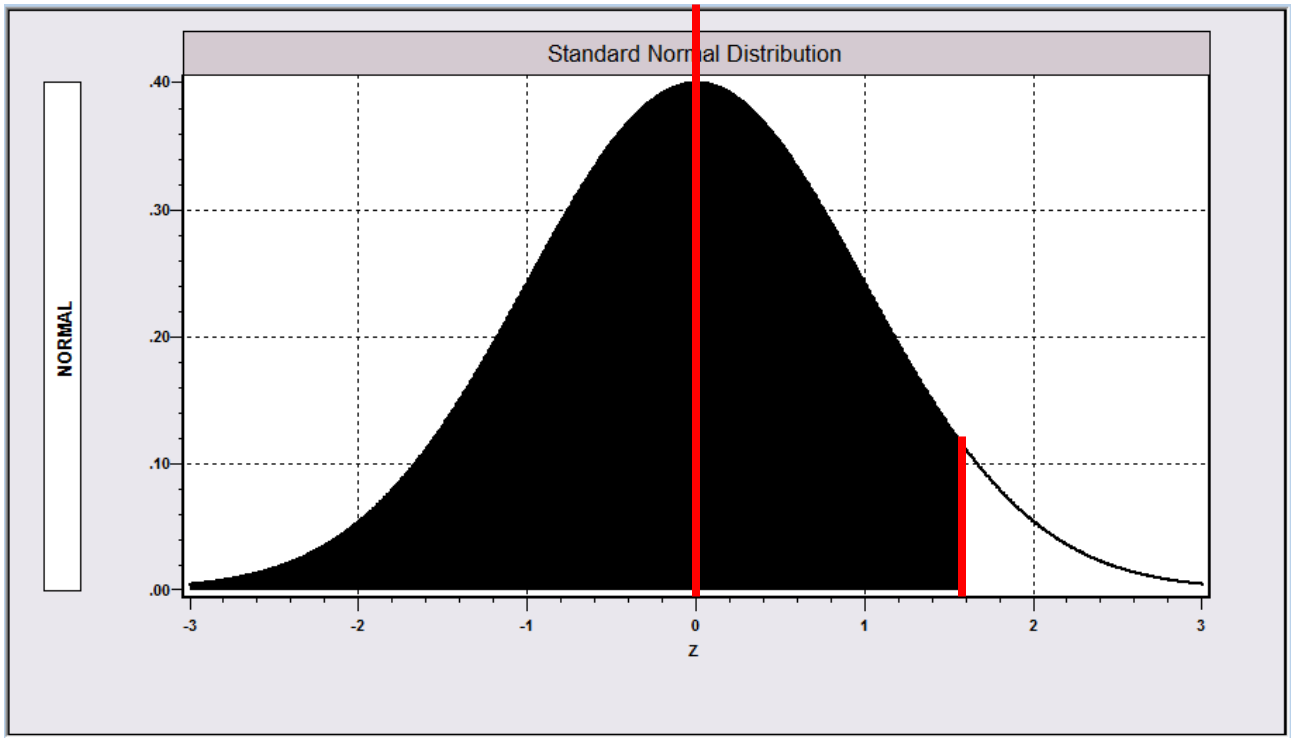
Standardized Normal – Compute Probabilities for $N[\mu, \sigma^2]$ by using N[0,1]

- $X \sim N[\mu, \sigma^2]$
- $\text{Prob}[X \leq a] = F(a)$
- $\text{Prob}[X \leq a] = \text{Prob}[(X - \mu)/\sigma \leq (a - \mu)/\sigma]$
  $= \text{Prob}[\text{Normal}(0,1) \leq (a - \mu)/\sigma]$

Only Half the Table Is Needed

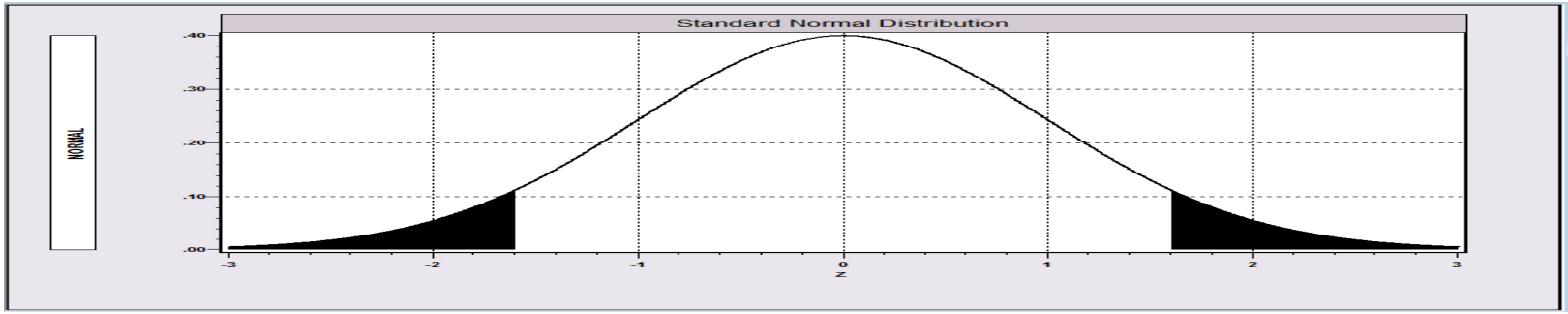


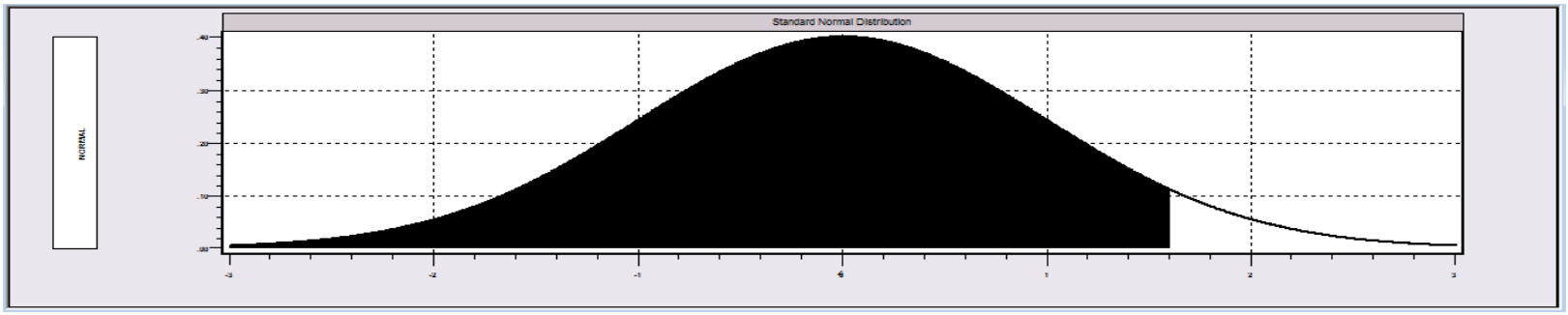
Only Half the Table Is Needed



The area left of 1.60 is exactly 0.5 plus the area between 0.0 and 1.60.

Areas Left of Negative Z

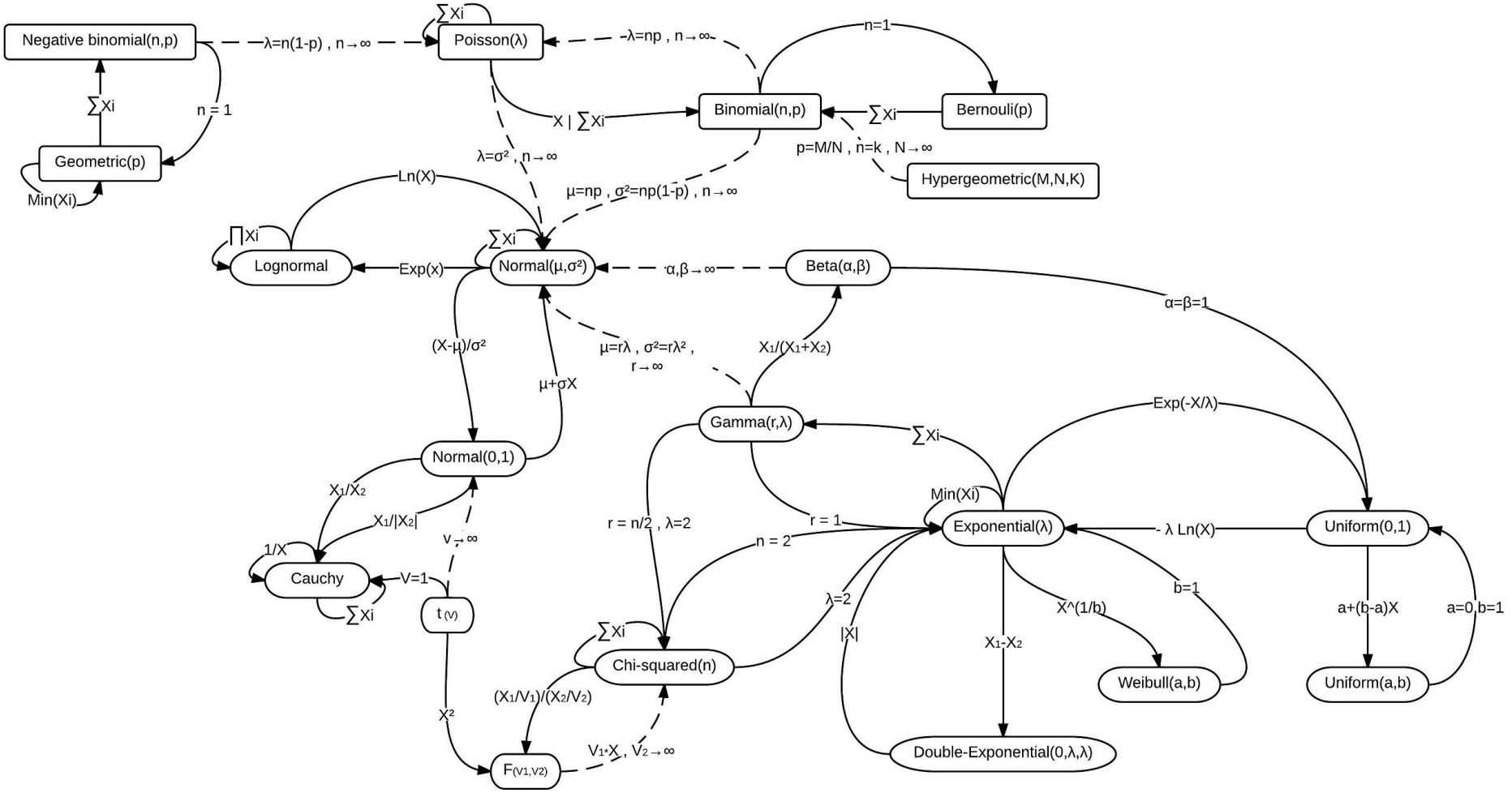




Area right of +1.6 equals 1 – area to the left of +1.6.

Computing Probabilities by Standardizing: Example

$$
\begin{aligned}
&P\left[4.5 \le x \le 8 \mid \mu = 3.5, \sigma = 2.0\right] \\
&= P\left[\frac{4.5-\mu}{\sigma} \le \frac{x-\mu}{\sigma} \le \frac{8-\mu}{\sigma}\right] \\
&= P\left[\frac{4.5-3.5}{2.0} \le \frac{x-3.5}{2.0} \le \frac{8-3.5}{2.0}\right] \\
&= P[0.5 \le z \le 2.25] \\
&= P[z \le 2.25] - P[z \le 0.5] \\
&= 0.9878 - 0.6915 \\
&= 0.2963
\end{aligned}
$$



By Ehsan.azhdari - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=22628518

BIA 652, Multivariate Data Analytics

Mean and Variance of Random Variables

Expected Value of a Random Variable

- Weighted average of the values taken by the variable

Discrete

$$E[X] = \sum_{\text{all values taken by X}} x \, \text{Prob}(X = x)$$

Continuous

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

(Density equals zero outside the range of x.)

Discrete Uniform

- X = 1,2,…,J
- Prob(X = x) = 1/J
- E[X] = 1/J + 2/J + … + J/J
  = J(J+1)/2 * 1/J
  = (J+1)/2
- Expected toss of a die = 3.5 (J=6)

Poisson (λ)

$$E[X] = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^{x}}{x!}$$

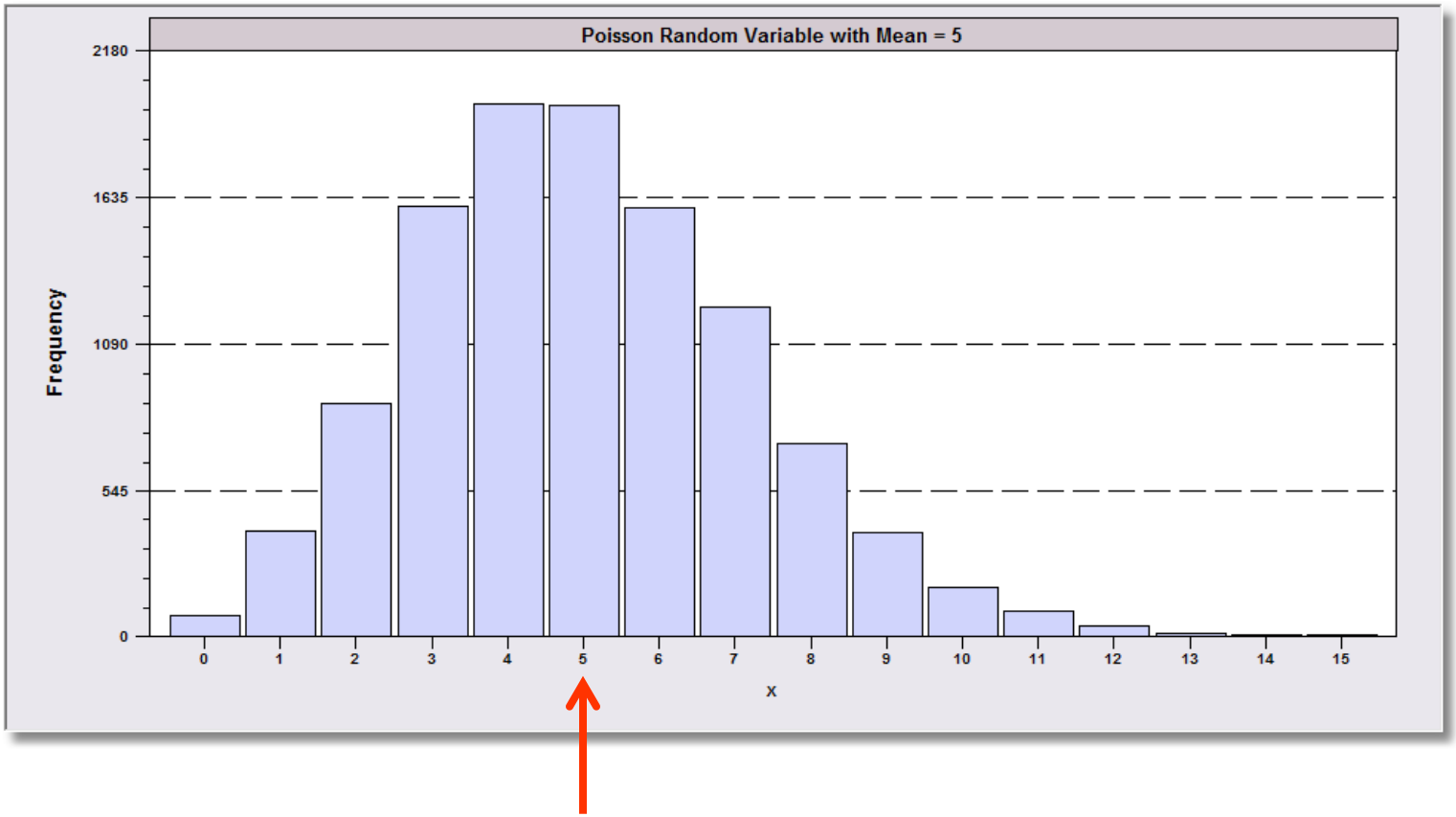
$$= \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^{x}}{x!} \text{ (drop zero term)}$$

$$= \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \left(\text{factor out } \lambda \text{ and } \frac{x}{x!} = \frac{1}{(x-1)!} \right)$$

$$= \lambda \sum_{z=0}^{\infty} \frac{e^{-\lambda}\lambda^{z}}{z!} \text{ (let } z = x\text{-}1\text{; } z \text{ goes from 0 to } \infty)$$

$$= \lambda \text{ (probabilities sum to 1)}$$

Poisson (5)



BIA 652, Multivariate Data Analytics

Continuous Random Variable

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx, \text{ 'support of x'} = \{x : f(x) > 0\}$$

$$\text{Continuous uniform: } f(x) = \frac{1}{b-a} I(x \in [a,b])$$

$$E[x] = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a}\left(\frac{x^2}{2} \Big|_a^b \right)$$

$$= \frac{b^2 - a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2} \quad \text{(the midpoint)}$$

Expected Value of a Function of X

- Y=g(X)
- **E[y] is generally not equal to g(E[X]) if g(X) is not linear**

THEOREM A

Suppose that $Y = g(X)$.

a. If X is discrete with frequency function $p(x)$, then

$$E(Y) = \sum_x g(x)p(x)$$

provided that $\sum |g(x)|p(x) < \infty$.

b. If X is continuous with density function $f(x)$, then

$$E(Y) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$$

provided that $\int |g(x)|f(x)\,dx < \infty$.

Expected Value of a Linear Translation

- $Z = aX+b$
- $E[Z] = aE[X] + b$
- Example: toss of a die

Properties There are some properties of the expectation. For random variables $X_1, X_2, ...,$ and X_n,

Addition Rule	$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$	however dependent $X_1, X_2, ..., X_n$ are;
Multiplication Rule	$E(X_1 X_2) = E(X_1)E(X_2)$	if X_1 and X_2 are independent;
Scaling & Shifting	$E(aX + b) = aEX + b$	for constants a and b.

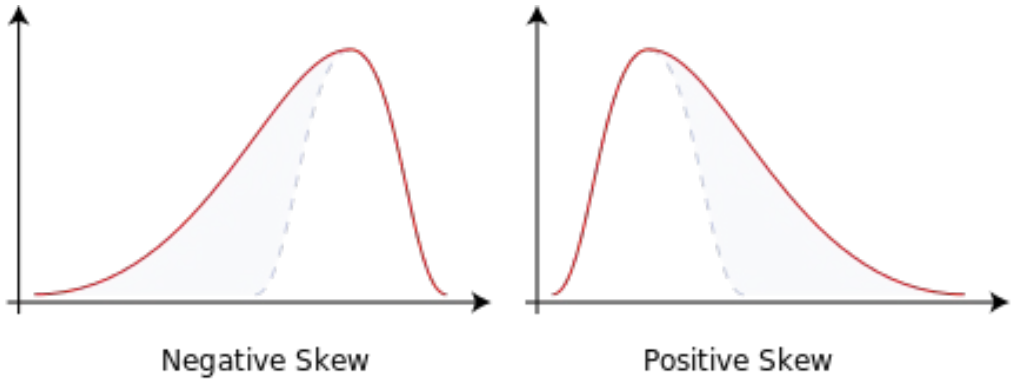
E(Powers of X) - Moments

- Moment = $E[X^k]$ for positive integer k
  - Raw moment: $E[X^k]$
  - Central moment: $E[(X - E[X])^k]$
  - Standardized moment: $E[((X - E[X])/\sigma)^k]$

- The first moment is the expected value, the second central moment is the variance

The skewness of a random variable X is the third standardized moment $\tilde{\mu}_3$, defined as:

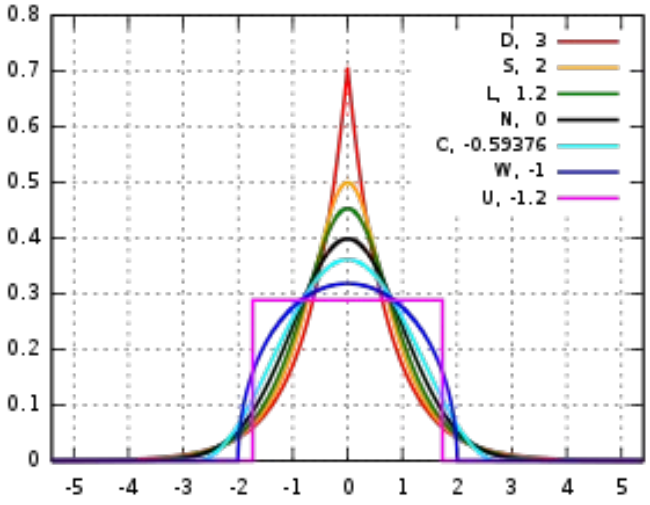
$$\tilde{\mu}_3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}[(X-\mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$



The kurtosis is the fourth standardized moment, defined as

$$\mathrm{Kurt}[X] = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4},$$

Source: Wikipedia



Variance

Variance = $E[(X - E[X])^2]$

Discrete

$$Var[X] = \sum_x (x - \mu)^2 \, Prob(X = x)$$

Continuous

$$Var[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Standard deviation = square root of variance

Variance of rolling a die?

Variance of Uniform[0, 1]?

Variance of a Translation: Y = a + bX

- Var[a] = 0
- Var[bX] = b^2Var[X]
- Standard deviation of Y = |b| Std.Dev.(X)

Shortcut

- $\text{Var}[X] = E[X^2] - \{E[X]\}^2$

Uniform (0,1)

$$E[X] = \frac{1}{2}$$

$$E[X^2] = \int_0^1 x^2 1 dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

$$\text{Var}[X] = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$

Using Excel & Python to compute probabilities

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

Thank you!

Prof. Feng Mai
School of Business

For academic use only.