ALY 6050[21598]

Introduction to Enterprise Analytics

Northeastern

M1: Review of Probability Distributions

Richard He
College of Professional Studies
Monday, Winter A, 2022

Prerequisites

- Excel
- R programming experience
- Completion of ALY6010 & ALY6015

Administrative Notes

- Academic Integrity
  - Cheating
  - Fabrication
  - Plagiarism
  - Unauthorized Collaboration
  - …

https://osccr.sites.northeastern.edu/academic-integrity-policy/

Administrative Notes

- Discussion:
  - One primary response(250 words) by **Thursday**
  - Two secondary responses(100 words) by **Sunday**
  - Last discussion is due by Friday of week 6
- Weekly Projects: due on **Sunday**
  - Last project is due by Friday of week 6
- TA: Sashank Yakkali
  yakkali.s@northeastern.edu

Grade Breakdown

Assignment	Grade	Weight in Course Grade
6 Weekly Discussions	240 points (40 points each)	24%
5 Weekly Projects	500 points (100 points each)	50%
Final Assessment	100 points	10%
Week 6 Final Project	160 points	16%
Total:	1000 points	100%

Topics Covered in this Course

- Review of Probability Distributions
- Simulation
- Forecasting & Regression
- Decision Modeling
- Optimization I
- Optimization II

Learning Objectives

- Discrete distributions:
  - Bernoulli, binomial, hypergeometric, and Poisson
- Continuous distributions:
  - uniform, triangular, normal, exponential, beta, gamma, log-normal, Weibull
- Goodness-of-fit test

Lecture Part I Recap

Probability vs. Statistics

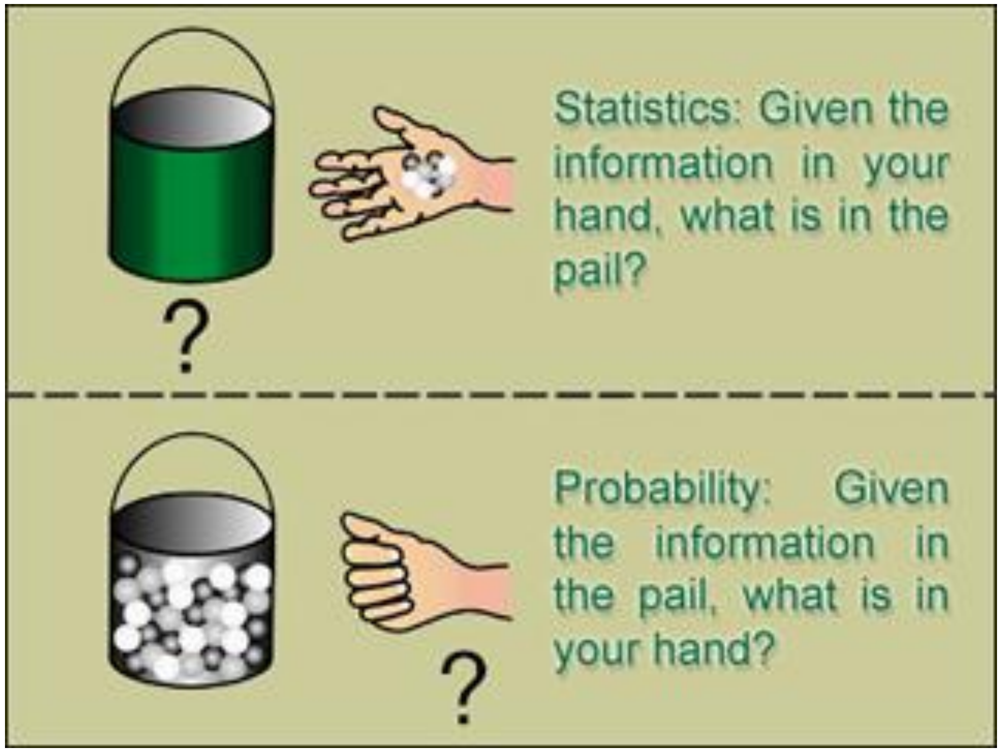


Image by MIT OpenCourseWare. Based on Gilbert, Norma. *Statistics.* W.B. Saunders Co., 1976.

Descriptive Statistics

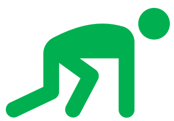
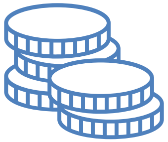
- Measures of central tendency
  - mean, median, mode
- Measures of dispersion or variation
  - range, variance, standard deviation
- Measures of position
  - percentile, quartile
- Measures of frequency
  - count, percent

Random Variable(RV)

- A **variable** is an attribute of an object of study
  - Categorical
  - Numerical
- A random variable is associated with **uncertainty**

Discrete vs. Continuous Variable

- Discrete
  - Only certain values
  - **Countable** & Finite

  - Number of people in a race, coin toss

- Continuous
  - **Any value** on an interval
  - **Measurable** & Infinite

  - Time to run a race, amount of snow in winter

Expectation & Variance

$$E[X] = \mu_x = \sum_{i=1}^{n} x_i p_i = \bar{X}$$

$$Var[X] = \sigma_x^2 = \sum_{i=1}^{n} (x_i - \mu_x)^2 p_i$$

Variance

$$Var[X] = \sigma_x^2 = \sum_{i=1}^{n} (x_i - \mu_x)^2 p_i$$

$$= \sum_{i=1}^{n} {x_i}^2 p_i - 2\,\mu_x x_i p_i + {\mu_x}^2 p_i$$

$$= E[X^2] - 2\mu_x \sum_{i=1}^{n} x_i p_i + {\mu_x}^2$$

$$= \boldsymbol{E[X^2] - {\mu_x}^2}$$

Covariance

$$
\begin{aligned}
Cov[X,Y] &\stackrel{\text{def}}{=} E[(X-\bar{X})\,(Y-\bar{Y})] \\
&= E[XY] - E[X]E[Y] \\
&= \boldsymbol{E[XY] - \mu_x\,\mu_y}
\end{aligned}
$$

$$Cov[X,X] = Var(X)$$

Correlation

$$\boldsymbol{Cov[U,V]} = Cov\left[\frac{X-\mu_x}{\sigma_x}, \frac{Y-\mu_y}{\sigma_y}\right]$$

$$= E\left[\frac{X-\mu_x}{\sigma_x}\frac{Y-\mu_y}{\sigma_y}\right] - E\left[\frac{X-\mu_x}{\sigma_x}\right]E\left[\frac{Y-\mu_y}{\sigma_y}\right]$$

$$= \frac{E[XY]-\mu_x\mu_y}{\sigma_x\sigma_y} - \frac{E[X]E[Y]-\mu_x\mu_y}{\sigma_x\sigma_y}$$

$$= \frac{E[XY]-E[X]E[Y]}{\sigma_x\sigma_y} = \boldsymbol{\frac{Cov[X,Y]}{\sigma_x\sigma_y} = \rho_{xy}}$$

Markov's Inequality

$$P(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon}$$

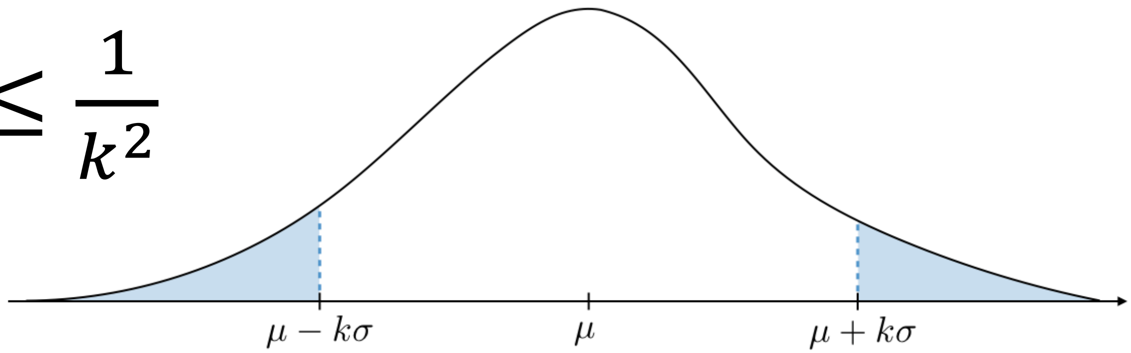
$$E[X] = \int_0^\infty xf(x)dx \geq \int_\varepsilon^\infty xf(x)dx$$

$$\geq \int_\varepsilon^\infty \varepsilon f(x)dx = \varepsilon \int_\varepsilon^\infty f(x)dx$$

$$= \varepsilon P(X \geq \varepsilon)$$

Chebyshev's Inequality

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$



$$P(|X-\mu| \geq k\sigma) = P((X-\mu)^2 \geq k^2\,\sigma^2)$$

$$\leq \frac{E[(X-\mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

Chebyshev's Inequality II

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

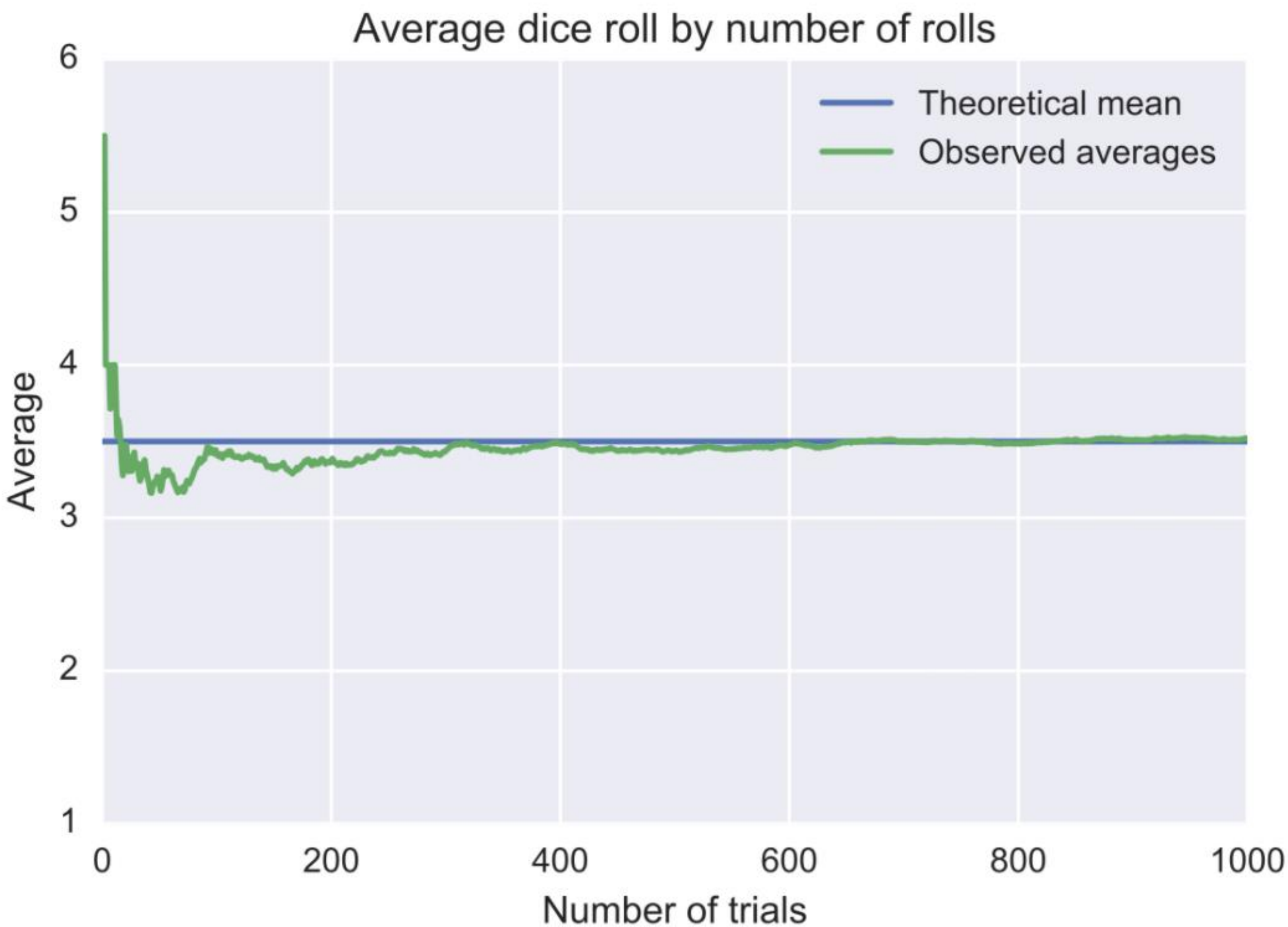
$$P(|X-\mu| \geq \varepsilon) = P((X-\mu)^2 \geq \varepsilon^2)$$

$$\leq \frac{E[(X-\mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$$

Law of Large Numbers (LLN)

- Sample mean **converges** to population mean as sample size n → infinity
- No guarantee when sample size is small

Illustration for LLN



https://en.wikipedia.org/wiki/Law_of_large_numbers

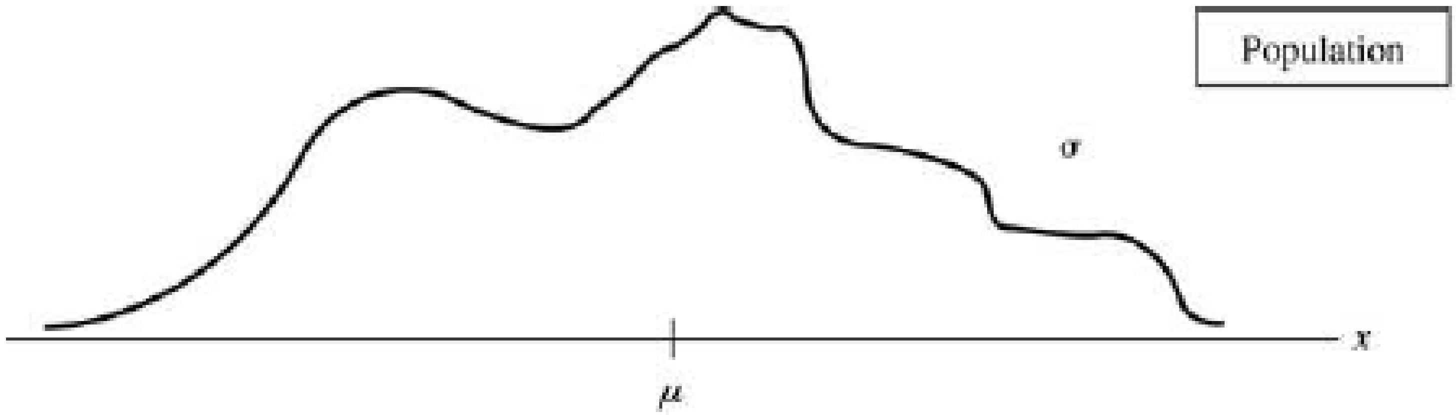
Central Limit Theorem (CLT)

- Regardless the underly population
- The distribution of sample means is approximately **normal**
  - mean: $\mu_{\bar{x}} = \mu$
  - standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Confidence Interval $100(1-\alpha)\%$:

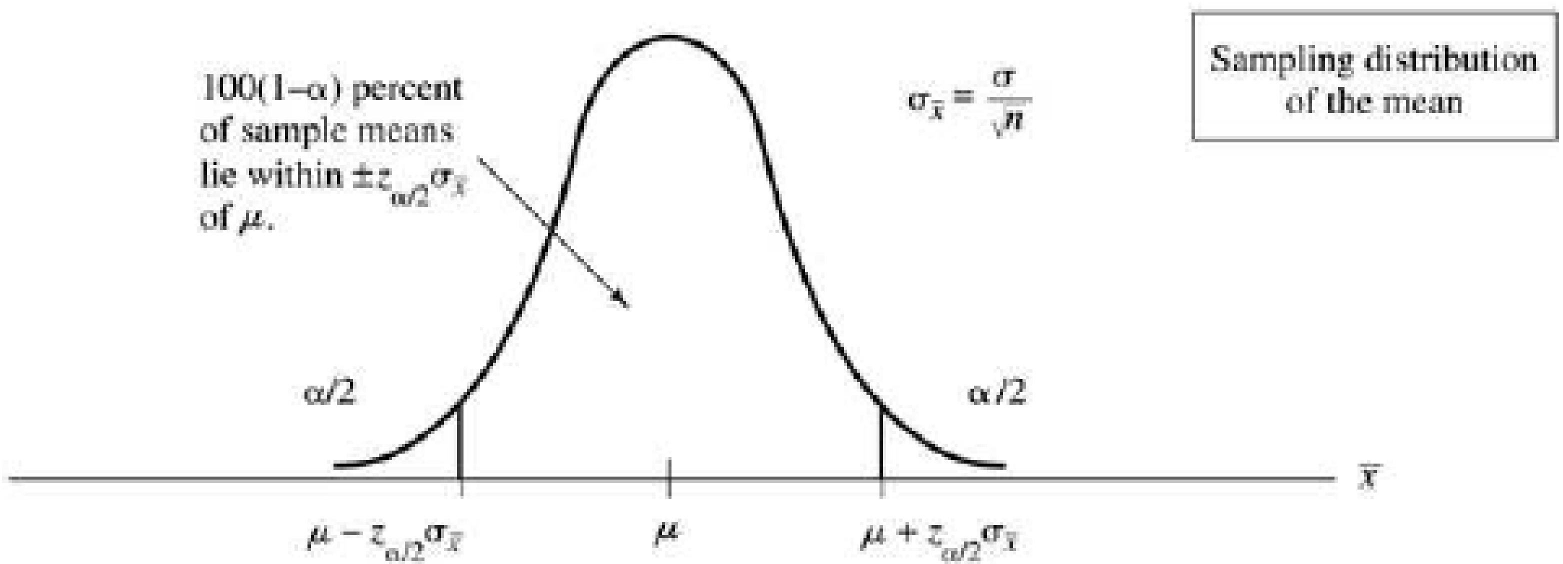
$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

```
> qnorm(0.025, lower.tail = FALSE)
[1] 1.959964
> qnorm(0.05, lower.tail = FALSE)
[1] 1.644854
```

Sampling Distribution of the Mean



$$z_{0.025} = 1.96 \text{ and } z_{0.05} = 1.645$$



Probability Distributions

Distribution	Abbreviation	Distribution	Abbreviation
Beta	beta	Logistic	logis
Binomial	binom	Multinomial	multinom
Cauchy	cauchy	Negative binomial	nbinom
Chi-squared (noncentral)	chisq	Normal	norm
Exponential	exp	Poisson	pois
F	f	Wilcoxon signed rank	signrank
Gamma	gamma	T	t
Geometric	geom	Uniform	unif
Hypergeometric	hyper	Weibull	weibull
Lognormal	lnorm	Wilcoxon rank sum	wilcox

R Probability Functions

- [d]ensity
- [p]robability
- [q]uantile
- [r]andom

R Probability Functions - Normal

- dnorm()
- pnorm()
- qnorm()
- rnorm()

Bayes' Theorem

H: Hypothesis
D: Data or Event

Likelihood
How probable is the event given the hypothesis is true

Prior
How probable was the hypothesis before observing any event?

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)}$$

Posterior
How probable is the hypothesis given the observed event?

Marginal
How probable is the new event under all possible hypothesis?

$$P(D) = \sum P(D|H_i)P(H_i)$$

Food for Thought

- Q1. A special party invites family with **at least one son**. Bob's family has two children and gets invited. What's the probability that both children are boys?

- Q2. Bob's family has two children. If you see one of his children is boy, what's the probability that both children are boys?

Lecture Part II
Discrete Probability Distributions

Discrete Probability Distributions

- Bernoulli Distribution
- Binomial Distribution
- Multinomial Distribution
- Hypergeometric Distribution
- Geometric Distribution
- Poisson Distribution

Bernoulli Distribution

- Two possible outcomes (**mutually exclusive**)
  - p: success
  - q: failure (q = 1 - p)
- One trial

$$X \sim Bernoulli(p), 0 \leq p \leq 1$$

$$p(x) = p^x(1-p)^{1-x}, x = 0 \; or \; 1$$

Ex for Bernoulli

- 1000 tickets are sold at $1 each
- Winner gets $750
- What's the expected value of the gain if you purchase one ticket?

	Win	Lose
Gain X	$749	−$1
Probability $P(X)$	$\frac{1}{1000}$	$\frac{999}{1000}$

$$E(X) = \$749 * \frac{1}{1000} - \$1 * \frac{999}{1000} = -\$0.25$$

Binomial Distribution

- Two possible outcomes (**mutually exclusive**)
- n trials
  - Independent
  - Identical

$$X \sim Binomial(n, p), 0 \leq p \leq 1$$

$$p(x) = C_n^x p^x (1-p)^{n-x}$$

Ex for Binomial

- One out of five Americans has visited a doctor in any given month
- Randomly selected 10 people
- Find the probability that exact 3 people have visited a doctor last month.

$$p = \frac{1}{5}, q = \frac{4}{5}, n = 10, x = 3$$

Ex for Binomial

$$p(x) = C_n^x p^x (1-p)^{n-x}$$

$$p(3) = C_{10}^3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7 \approx 0.201$$

```
> dbinom(3,10,0.2)
[1] 0.2013266
```

Multinomial Distribution

- More than two possible outcomes
- n trials
  - Independent
  - Identical

$$X \sim Multinomial(n, p)$$

$$p(x) = \frac{n!}{x_1! \cdot x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

$$\begin{aligned} &where \quad x_1 + x_2 + \cdots + x_k = n \\ &and \quad p_1 + p_2 + \cdots + p_k = 1 \end{aligned}$$

Ex for Multinomial

- 65% use herbicides for commercial purposes
- 27% for agricultural purposes
- 8% for home and garden purposes
- Find the probability that 3 used them for commercial purposes, 1 for agriculture, and 1 for home or garden purposes.

$$n = 5, where\ \ x_1 = 3, x_2 = 1, x_3 = 1$$
$$p_1 = 0.65, p_2 = 0.27, p_3 = 0.08$$

Ex for Multinomial

$$p(x) = \frac{n!}{x_1! \cdot x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

$$p(x) = \frac{5!}{3!1!1!} (0.65)^3 \, (0.27)^1 \, (0.08)^1$$
$$\approx 0.119$$

```
> dmultinom(x=c(3,1,1), size=5, prob=c(0.65,0.27,0.08))
[1] 0.118638
```

Hypergeometric Distribution

- Similar to binomial
- Sampling **without replacement**
- Applications: quality control

$$X \sim Hyper(m, n, k)$$

$$p(x) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k}$$

Ex for Hypergeometric

- A lot of **12** compressor tanks with 3 defectives
- Three tanks are checked for leaks.
- If 1 or more of the 3 is defective → **rejected**
- Find the probability rejecting the lot

m = 3 defective tanks
n = 9 good tanks
k = 3 number of tank checked
x = 0 no defective in the 3 tanks checked

Ex for Hypergeometric

$$p(0) = \frac{C_m^x C_n^{k-x}}{C_{m+n}^k} = \frac{C_3^0 C_9^3}{C_{3+9}^3} \approx 0.382$$

$$p(rejection) = 1 - p(0) = 0.618$$

```
> dhyper(x=0, m=3, n=9, k=3)
[1] 0.3818182
> 1-dhyper(x=0, m=3, n=9, k=3)
[1] 0.6181818
```

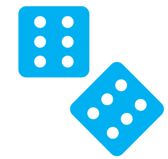
Geometric Distribution

- Two possible outcomes (**mutually exclusive**)
- n trials **until a success has been met**
  - Independent
  - Identical

$$X \sim Geometric(p), 0 \leq p \leq 1$$

$$p(n) = p(1-p)^{n-1}$$

Ex for Geometric



- Find the probability of getting the first 2 on the 3rd roll of a die.

$$p(not\ 2\ \&\ not\ 2\ \&\ 2) = \frac{5}{6}\frac{5}{6}\frac{1}{6} = \frac{25}{216}$$

$$p(3) = p(1-p)^{n-1} = \frac{1}{6}\left(\frac{5}{6}\right)^{3-1} = \frac{25}{216}$$

```
> dgeom(x=2, prob=1/6)
[1] 0.1157407
> 25/216
[1] 0.1157407
```

Poisson Distribution

- Modeling the frequency of **rare events**.
- The occurrences are **random** and **independent**
- The average number of occurrences over an **interval** is known

$$X \sim Poisson(\lambda), \qquad \lambda > 0$$

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0,1, \ldots$$

Expectation of Poisson Distribution

$$E[X] = \lambda$$

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0,1, \ldots$$

$$E[X] = \sum_{i=1}^{n} x_i p_i$$

Ex for Poisson

- 200 typos in a 400-page manuscript.
- Find the probability of a given page having exactly **zero** errors.

```
> exp(1)^(-0.5)
[1] 0.6065307
> dpois(0, 0.5)
[1] 0.6065307
```

$$\lambda = \frac{200}{400} = 0.5$$

$$p(0) = e^{-\lambda}\frac{\lambda^x}{x!} = 2.7183^{-0.5} = 0.61$$

Lecture Part III Continuous Probability Distributions

Continuous Distributions

- Normal Distribution
- Log-normal Distribution
- Exponential Distribution
- Beta Distribution
- Gamma Distribution
- Weibull distribution

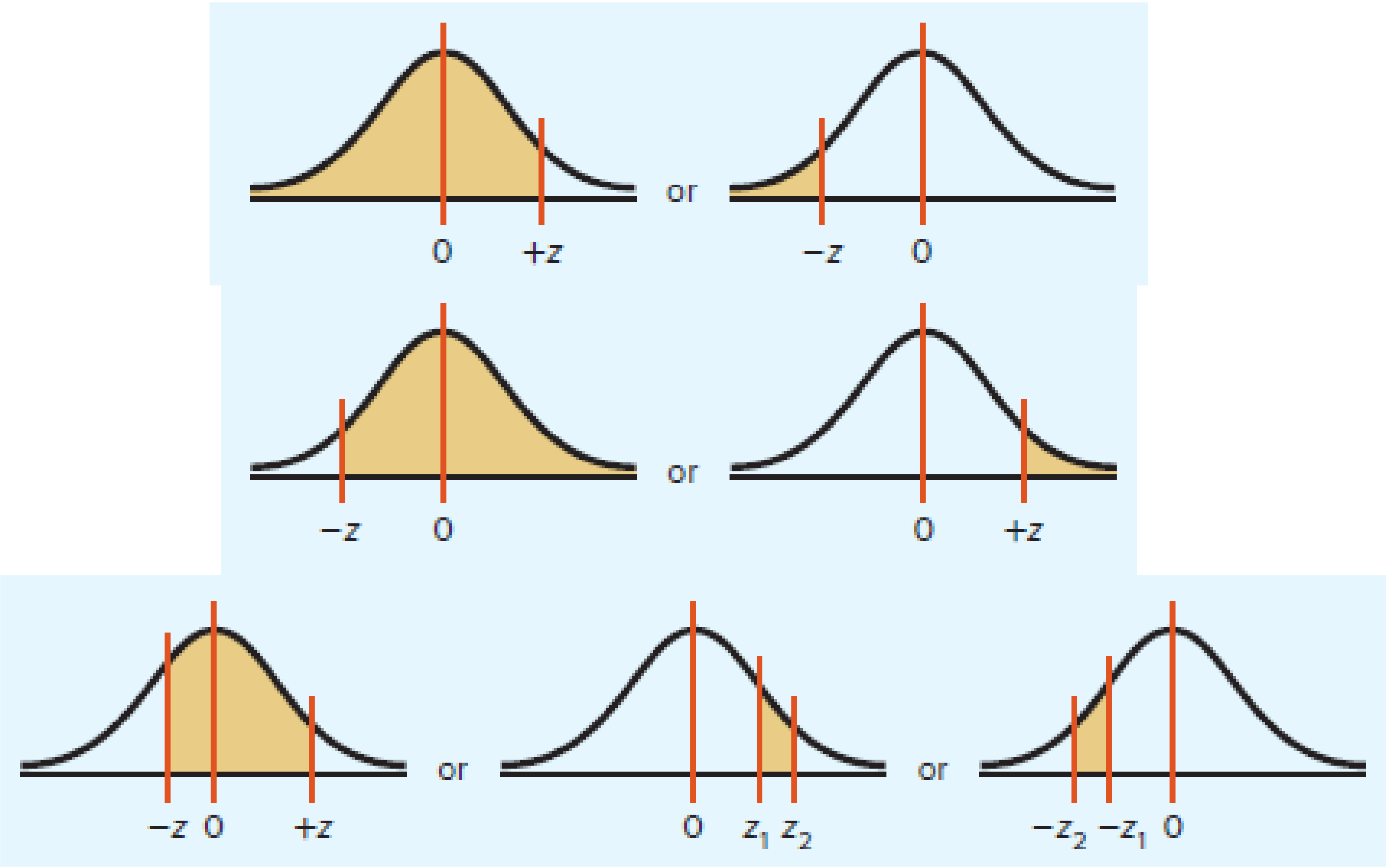
Normal Distribution

- Gaussian distribution or Bell-shaped curve
- The mean, median, and mode are **equal** and at center
- The area under the curve is 1
- The curve never touches the x-axis

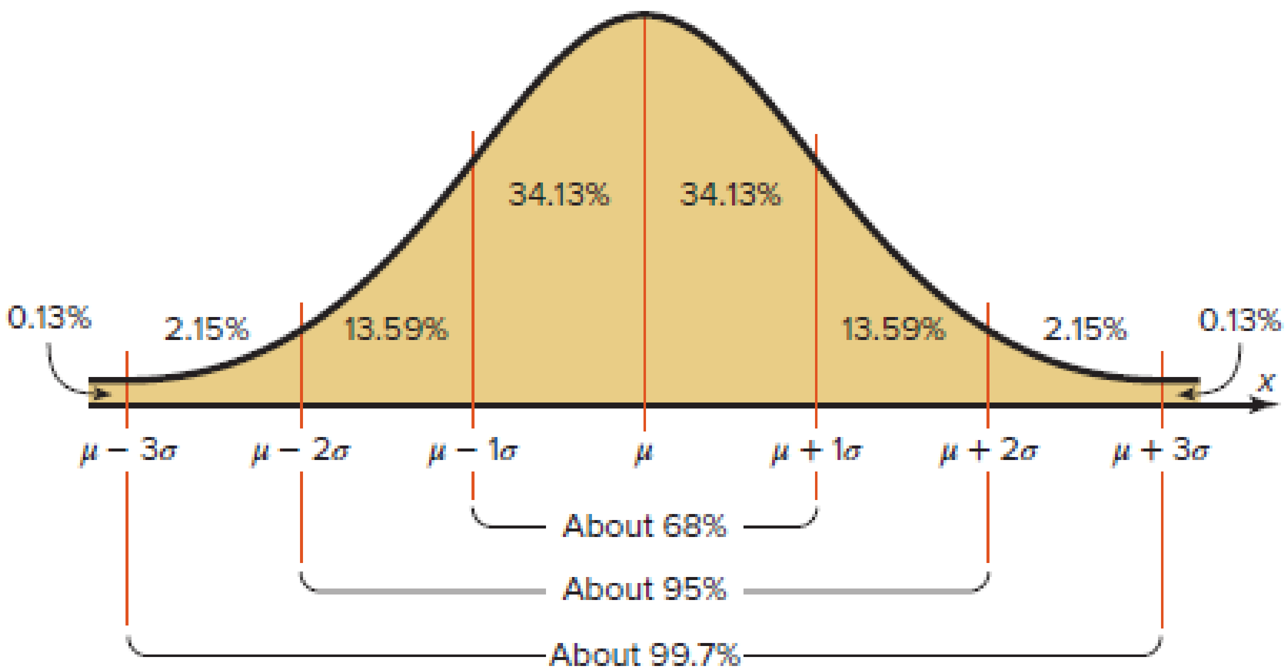
$$X \sim Normal(\mu, \sigma^2)$$

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

Calculate the Probability



Standard Normal Distribution



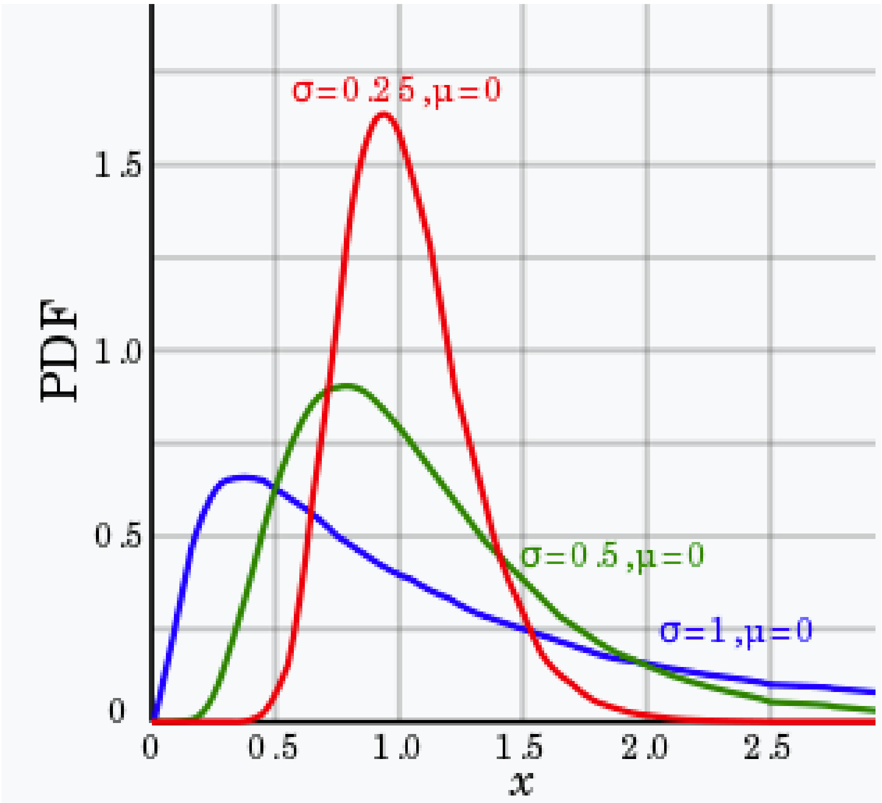

```
> pnorm(1)-pnorm(-1)
[1] 0.6826895
> pnorm(2)-pnorm(-2)
[1] 0.9544997
> pnorm(3)-pnorm(-3)
[1] 0.9973002
```

Lognormal Distribution

- A variable X is lognormally distributed if $Y = \ln(X)$ is normally distributed
- Applications:
  - time to complete a task

$$X \sim Lognormal(\mu, \sigma^2)$$
$$\ln(X) \sim N(\mu, \sigma^2)$$

Visualizing Log-normal



Identical parameter μ but differing parameters σ

https://en.wikipedia.org/wiki/Log-normal_distribution

Exponential Distribution

- One parameter
  - $rate\ \lambda$

- Applications: modeling **decaying** phenomena

$$X \sim Exponential(\lambda)$$

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

Ex for Exponential

- The mean time to failure of a critical component of an engine is 8000 hours
  - $\lambda = 1/8000$
- Find the probability of failing before 5000 hours.

$$F(5000) = 0.465$$

```
> pexp(5000, rate=1/8000)
[1] 0.4647386
```

Beta Distribution

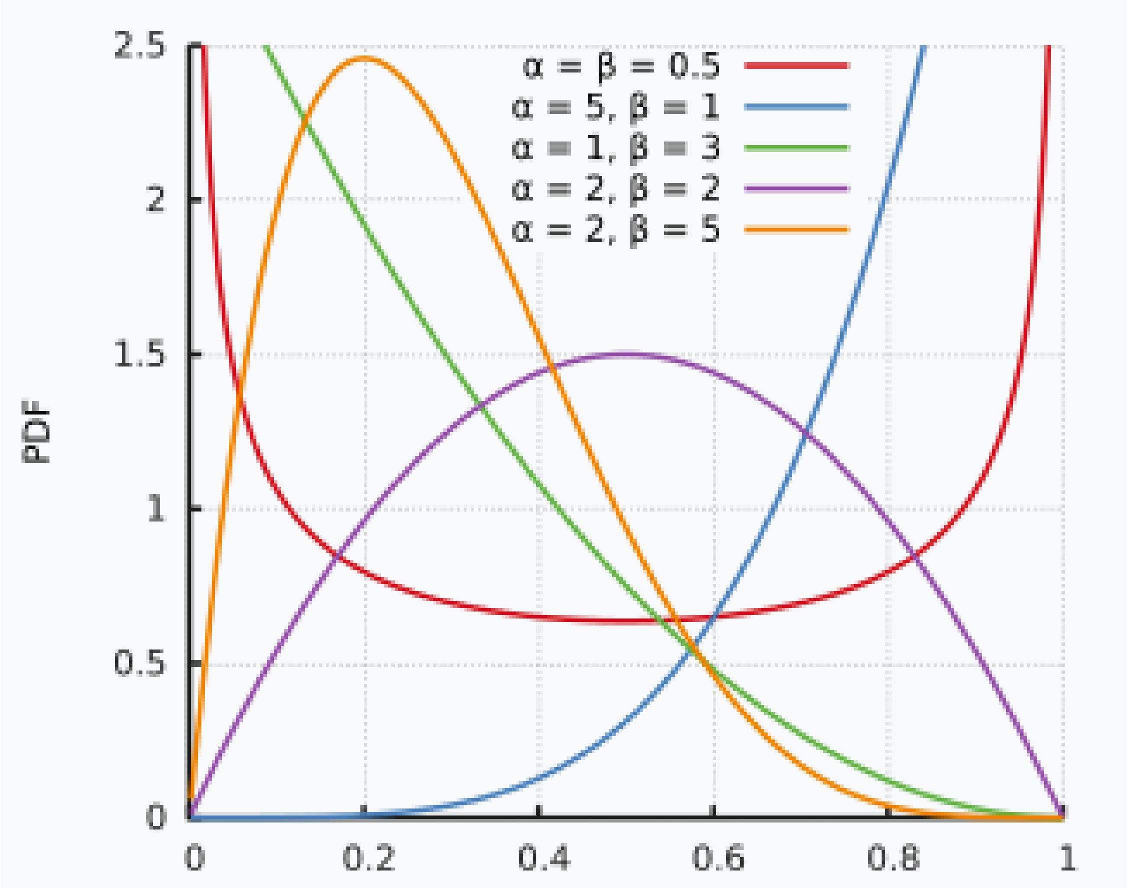
- Two parameters:
  - a shape parameter $\alpha > 0$
  - a shape parameter $\beta > 0$
- Applications:
  - Task cost and schedule modeling

$$X \sim Beta(\alpha, \beta)$$

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\mathrm{B}(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Visualizing Beta Distribution



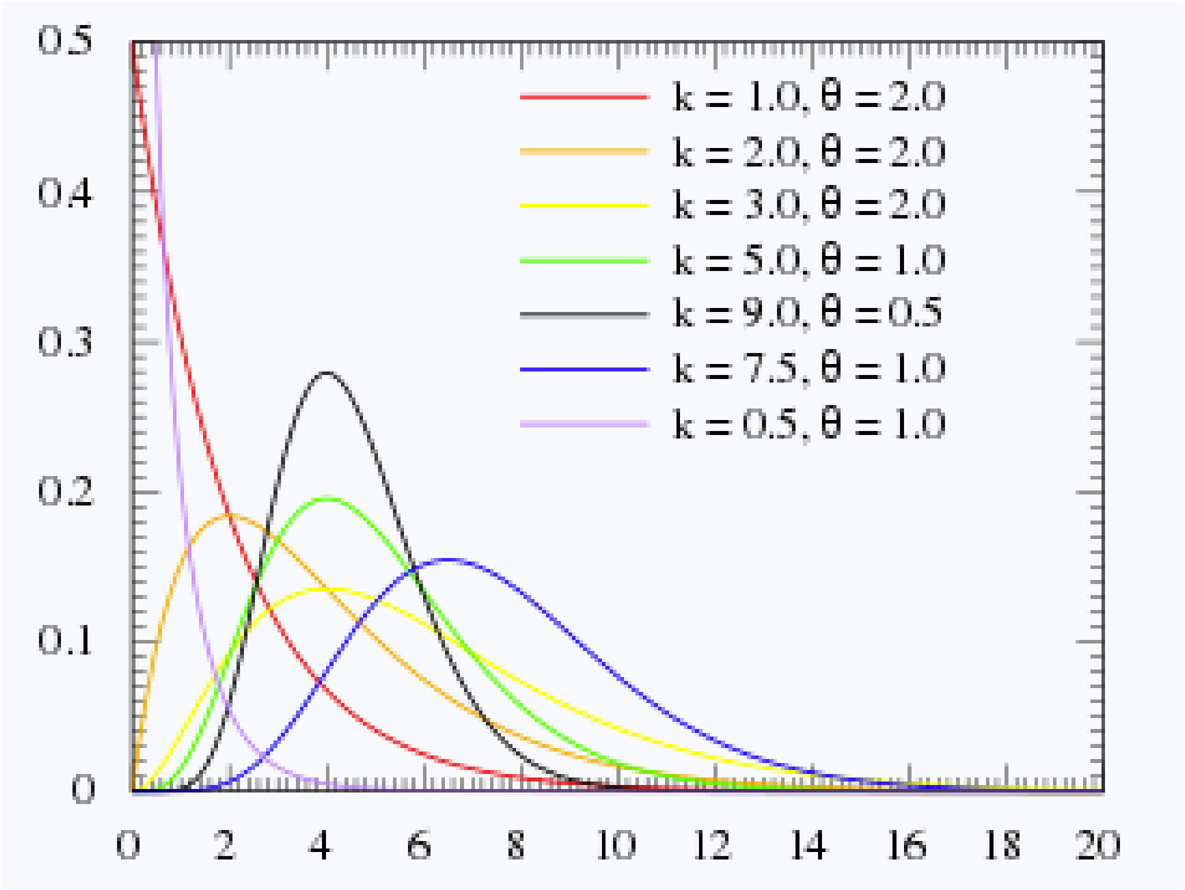
Gamma Distribution

- Two parameters:
  - a **shape** parameter $\alpha > 0$
  - a **rate** parameter $\beta > 0$
- Applications:
  - Size of loan defaults
  - Aggregate insurance claims
  - The amount of rainfall accumulated in a reservoir

$$X \sim Gamma(\alpha, \beta)$$

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Visualizing Gamma Distribution



$$k = \alpha, \theta = 1/\beta$$

Weibull Distribution

- Three parameters:
  - a location parameter $\gamma (-\infty < \gamma < +\infty)$
  - a scale parameter $\alpha > 0$
  - a shape parameter $\beta > 0$
- Applications: reliability analysis
  - time to failure of mechanical and electrical parts

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right], x \geq \gamma$$

Weibull Distribution

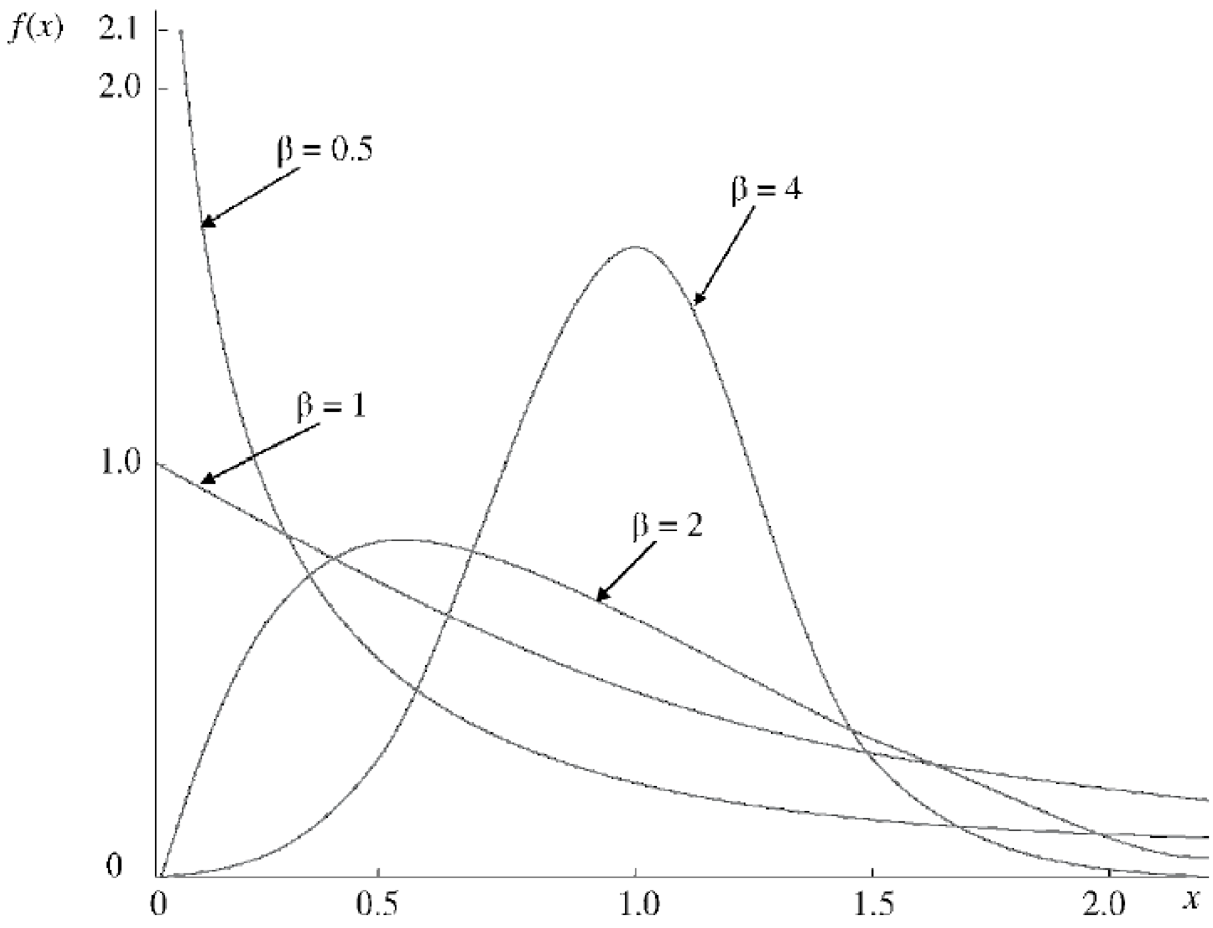
- Let $\gamma = 0$

$$f(x) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right], x \geq 0$$

- Let $\gamma = 0$ and $\beta = 1$

$$f(x) = \frac{1}{\alpha}\exp\left[-\frac{x}{\alpha}\right], x \geq 0$$

PDF for Weibull



$$\gamma = 0, \alpha = 1$$

Ex for Weibull

- The time to failure for a cathode ray tube (CRT) is modeled using Weibull distribution
  - location parameter $\gamma = 0$
  - scale parameter $\alpha = 200$ hours
  - shape parameter $\beta = \frac{1}{3}$

- What's the probability of a tube operating for at least 800 hours?

Ex for Weibull

$$F(x) = 1 - \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right], x \geq \gamma$$

$$F(x) = 1 - \exp\left[-\left(\frac{x}{200}\right)^{\frac{1}{3}}\right], x \geq 0$$

$$\begin{aligned} P(x > 800) &= 1 - P(x \leq 800) \\ &= 1 - \left\{1 - \exp\left[-\left(\frac{800}{200}\right)^{\frac{1}{3}}\right]\right\} \\ &= \exp\left[-(4)^{\frac{1}{3}}\right] \\ &\approx 0.204 \end{aligned}$$

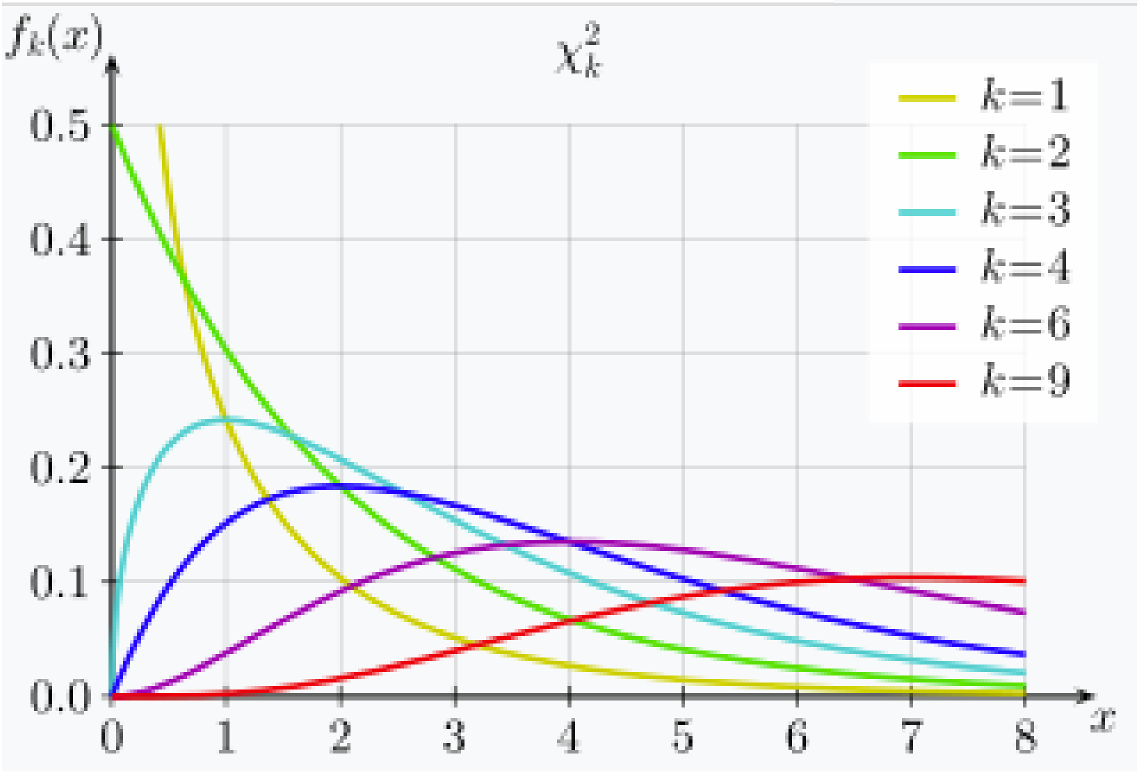
```
> pweibull(800, shape=1/3, scale=200, lower.tail=FALSE)
[1] 0.2044563
```

Lecture Part IV Goodness-of-fit Test

Chi-Square Distribution

- Based on degrees of freedom
- Positively skewed

$$\frac{1}{2^{k/2}\Gamma(k/2)}\,x^{k/2-1}e^{-x/2}$$



Chi-Square Goodness-of-Fit Test

- Degrees of freedom = no. of categories – 1
- O = observed frequency
- E = expected frequency

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Larceny thefts	Property crimes	Drug use	Driving under the Influence
38	50	28	44

Arrests for Crimes

Step 1. Hypotheses

H0: No difference in the number of arrests for each type of crime.

Ha: There is difference.

Step 2. Key parameters

degree of freedom = 4-1=3

$\alpha = 0.05$

critical value = 7.815

```
> qchisq(0.95, df=4-1)
[1] 7.814728
```

Arrests for Crimes

Step 3. Calculate test value (E=n/k=160/4=40)

	Larceny thefts	Property crimes	Drug use	Driving under the influence
Observed Expected	38 40	50 40	28 40	44 40

$$\chi^2 = \sum \frac{(0-E)^2}{E} = \frac{(38-40)^2}{40} + \frac{(50-40)^2}{40} + \frac{(28-40)^2}{40} + \frac{(44-40)^2}{40}$$

$$= 0.1 + 2.5 + 3.6 + 0.4$$

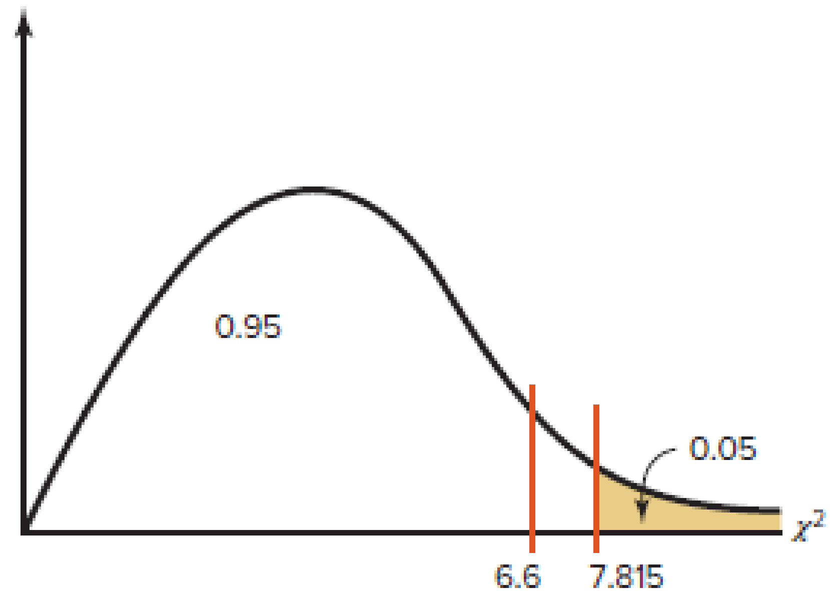
$$= 6.6$$

Arrests for Crimes

Step 4. Make decision

Since $6.6 < 7.815$

Do not reject H_0



Step 5. Summary

There is not enough evidence to reject the claim. Fail to reject.

Other Goodness-of-Fit Tests

- Anderson-Darling Test
  - ad.test {goftest}

- Kolmogorov-Smirnov Test
  - ks.test {stats}

Example: AD Test

```
> x <- rnorm(10, mean=2, sd=1)
> ad.test(x, "pnorm", mean=2, sd=1)

        Anderson-Darling test of goodness-of-fit
        Null hypothesis: Normal distribution
        with parameters mean = 2, sd = 1
        Parameters assumed to be fixed

data:  x
An = 1.3598, p-value = 0.2136
```

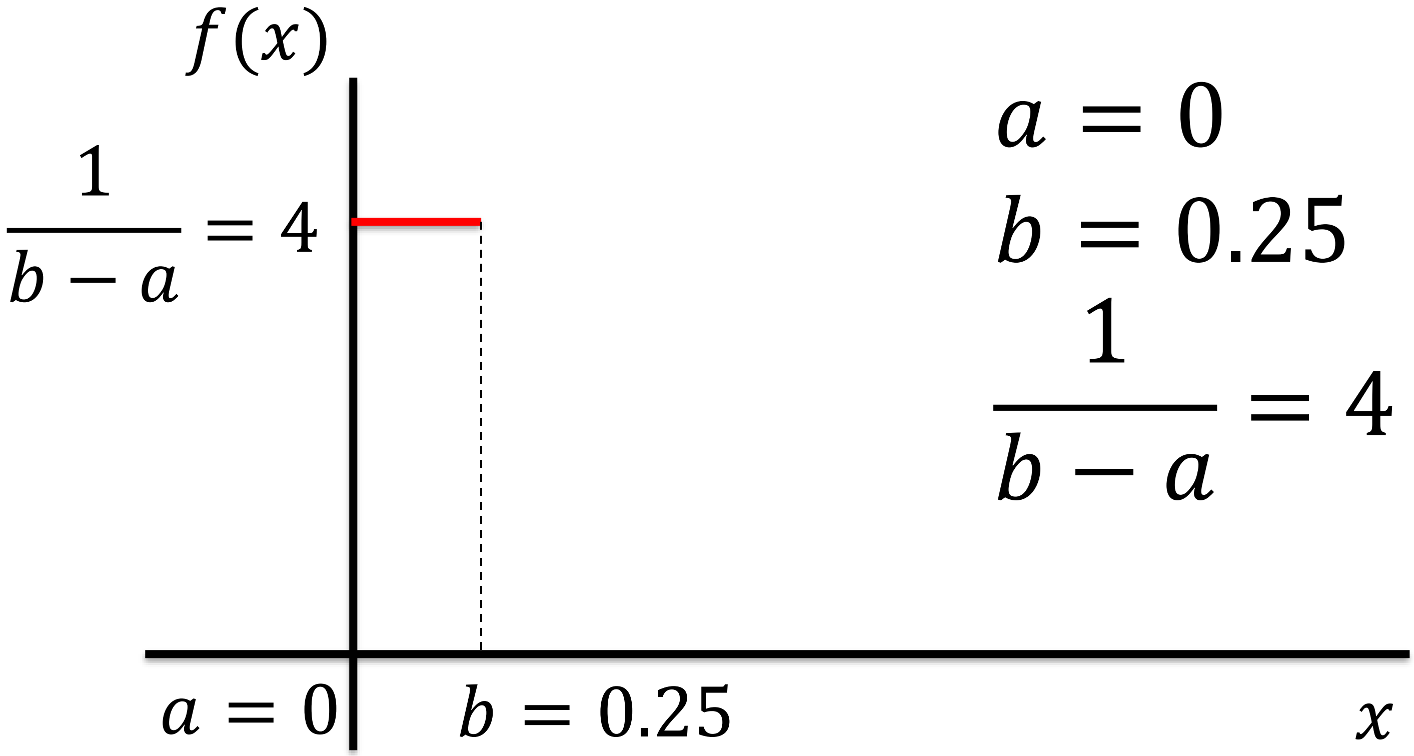
Example: KS Test

```
> x <- rnorm(50)
> y <- runif(30)
> ks.test(x, y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.48, p-value = 0.0002033
alternative hypothesis: two-sided
```

Questions?



$$a = 0$$
$$b = 0.25$$
$$\frac{1}{b-a} = 4$$

$$Probability = Area = 4 * 0.25 = 1$$