Introduction and Overview STAT 421, SP 2012

Prof. Prem K. Goel

Mon, Wed, Fri 3:30PM-4:48PM

Postle Hall 1180

Course Instructor

- Prof. Goel, Prem
- *E-mail*: goel.1@osu.edu
- *Office*: CH 204C (Cockins Hall)
- *Phone*: 614-292-8110
- Office Hours:
  - Tuesday 2:00 PM - 3:00 PM
  - Wednesday 1:30 PM - 2:30 PM
  - Friday 10:00 AM -11:00 AM
- Other times by appointment
- Course Website: www.stat.osu.edu/~goel/
- Relevant course material will be posted on this website by 8:00PM day before the class

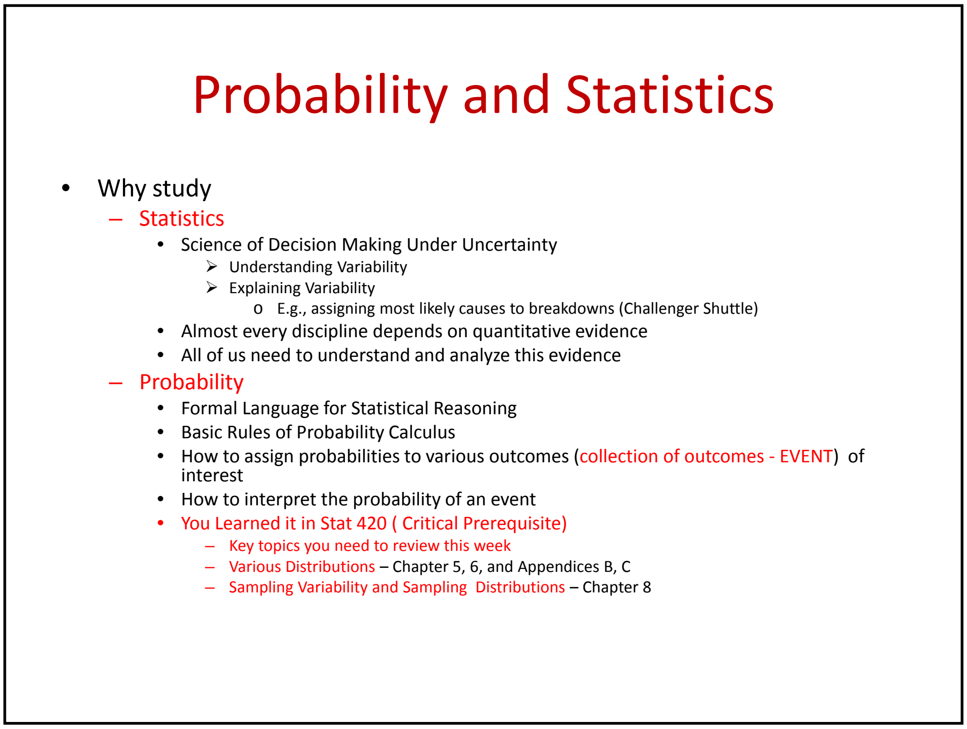
Recitation Instructors and Graders

- **Ms. Siyoen Kil**
- *Office: 304C Cockins Hall*
- *E-Mail*: kil.3@osu.edu
  - Th 11:30AM - 12:18PM
    Caldwell Lab 0277
  - Th 12:30PM – 01:18PM
    Dreese Lab 0369
- *Office Hours: TBD*

- **Ms. Lira Pi**
- *Office: 304F Cockins Hall*
- *E-mail*: pi.5@osu.edu
  - Th 10:30AM - 11:18AM
    Dreese Lab 0317
  - Th 12:30PM – 01:18PM
    Univ. Hall 0151
  - *Office Hours: Tuesdays, 10:00 AM -12:00 Noon*

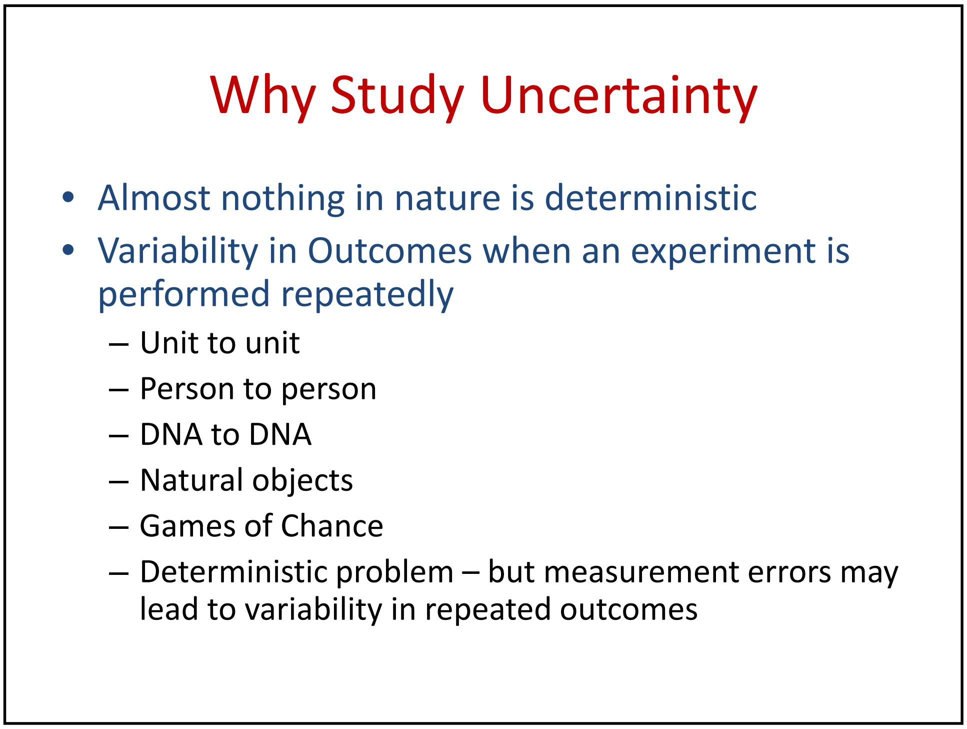
Student Information Needed

- Name: Last, First
- Signature
  - For matching with attendance sign-up sheet
- Major
- Math Courses Background



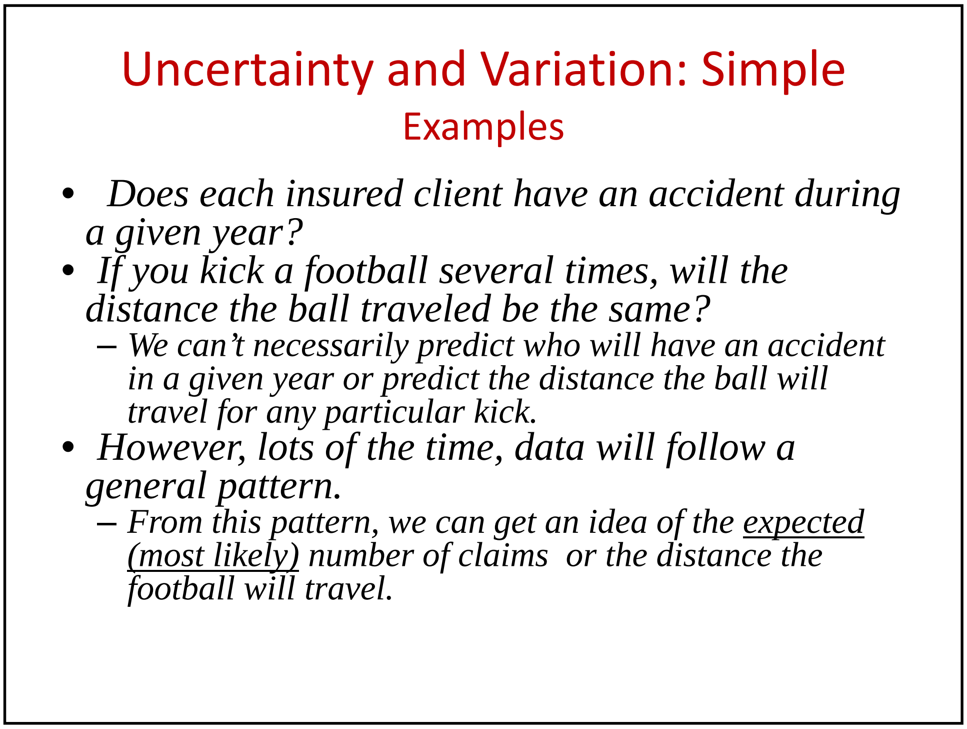
Probability and Statistics

- Why study
  - Statistics
    - Science of Decision Making Under Uncertainty
      - Understanding Variability
      - Explaining Variability
        - E.g., assigning most likely causes to breakdowns (Challenger Shuttle)
    - Almost every discipline depends on quantitative evidence
    - All of us need to understand and analyze this evidence
  - Probability
    - Formal Language for Statistical Reasoning
    - Basic Rules of Probability Calculus
    - How to assign probabilities to various outcomes (collection of outcomes - EVENT) of interest
    - How to interpret the probability of an event
    - You Learned it in Stat 420 (Critical Prerequisite)
      - Key topics you need to review this week
      - Various Distributions – Chapter 5, 6, and Appendices B, C
      - Sampling Variability and Sampling Distributions – Chapter 8



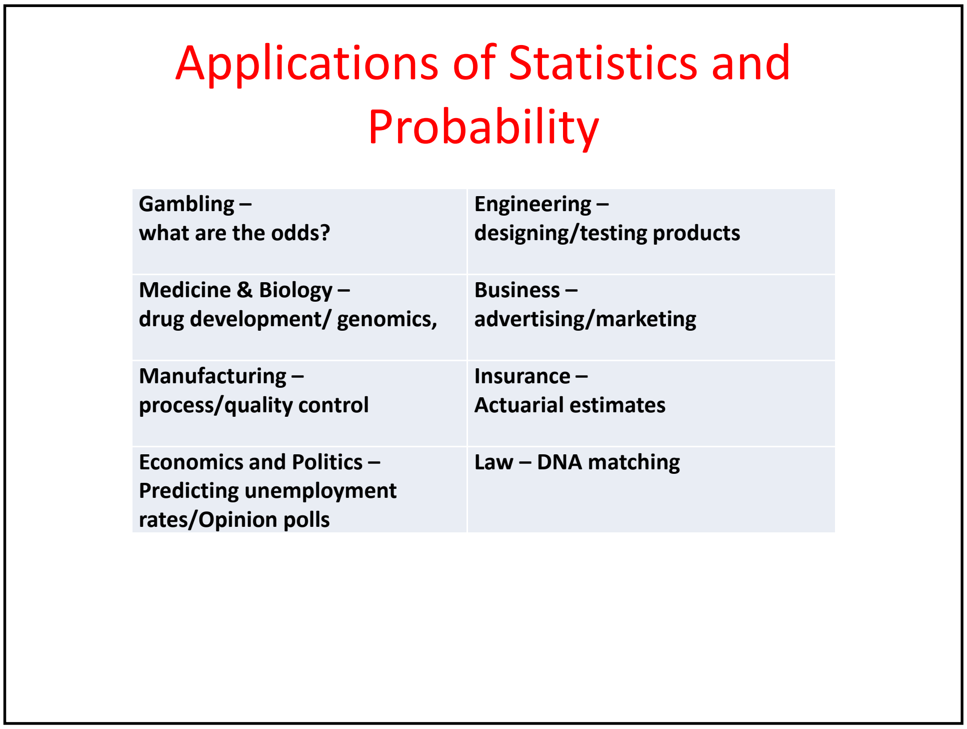
Why Study Uncertainty

- Almost nothing in nature is deterministic
- Variability in Outcomes when an experiment is performed repeatedly
  - Unit to unit
  - Person to person
  - DNA to DNA
  - Natural objects
  - Games of Chance
  - Deterministic problem – but measurement errors may lead to variability in repeated outcomes



Uncertainty and Variation: Simple Examples

- *Does each insured client have an accident during a given year?*
- *If you kick a football several times, will the distance the ball traveled be the same?*
  - *We can't necessarily predict who will have an accident in a given year or predict the distance the ball will travel for any particular kick.*
- *However, lots of the time, data will follow a general pattern.*
  - *From this pattern, we can get an idea of the expected (most likely) number of claims or the distance the football will travel.*

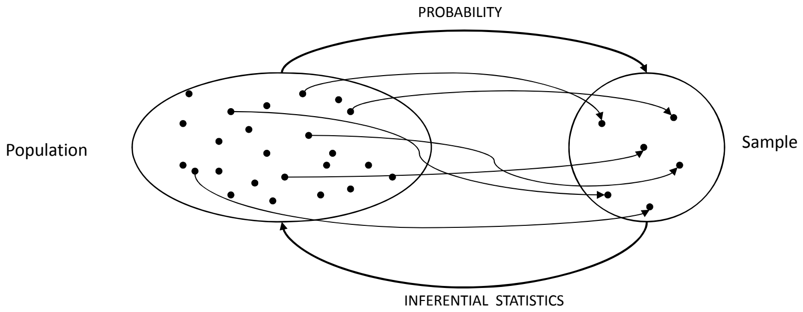


Applications of Statistics and Probability

Gambling – **what are the odds?**	**Engineering –** **designing/testing products**
Medicine & Biology – **drug development/ genomics,**	**Business –** **advertising/marketing**
Manufacturing – **process/quality control**	**Insurance –** **Actuarial estimates**
Economics and Politics – **Predicting unemployment rates/Opinion polls**	**Law – DNA matching**

Probability and Statistics

- The study of randomness and uncertainty
- "Chances", "odds", "likelihood", "expected", "probably", "on average", ...



Quick Review: Stat 420 Probability Concepts

Text Book:

Chapters 5, 6, and 8

Discrete Distributions

Distribution	$p_X(x)$	*Support*	$\mathrm{E}[X]$	$\mathbb{V}[X]$
$\mathrm{Bernoulli}(p)$	$p^x(1-p)^{1-x}$	$x=0,1$	p	$p(1-p)$
$\mathrm{Bin}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$x=0,\ldots,n$	np	$np(1-p)$
$\mathrm{Pois}(\lambda)$	$\frac{\lambda^x e^{-\lambda}}{x!}$	$x=0,1,2,\ldots$	λ	λ
Hypergeometric	$P(X=x)=\frac{\binom{r}{x}\times\binom{N-r}{n-x}}{\binom{N}{n}}$	$x=\max\{0,n-(N-r)\},\cdots,\min\{n,r\}$	$E(X)=np,\ \text{where } p=r/N.$	$Var(X)= fpc\cdot n\cdot p\left(1-p\right)$
Negative Binomial	$P(X=x)=\binom{x-1}{r-1}(1-p)^{x-r}\,p^r$	$x=r,\ r+1,\ \ldots$	$E(X)=\frac{r}{p}$	$V(X)=\frac{r(1-p)}{p^2}$

Before the Normal Distribution

- Normal Distribution – Bell Curve N(μ, σ^2)
  - Continuous distribution, defined on entire real line (allows positive density on negative numbers, even though it may be negligible)
  - Symmetric
- We may want to study a phenomena that has a skewed distribution? [For example taking all values in (0,∞)
  - Income and Consumption [Economics, Management]
  - Time until the next "hit" on a web page [Web Data Mining]
  - Response time to a stimulus [Psychology]
  - Pay-off for car-insurance policy [Insurance]
  - Time to Event [Insurance]
  - Lifetime of a component of a device [reliability studies]
  - Inter-arrival times of events [Transportation, Reliability, Queuing]

Gamma, Exponential and χ^2 Distributions

- A flexible "family" of distributions used to model these phenomena
  - The family can represent a large variety of shapes

Gamma Distribution: $X \sim \text{Gamma}(\alpha, \beta)$

Flexible, two-parameter family of distributions.

Special Cases:

Exponential Distribution: $X \sim \text{Exp}(\beta)$

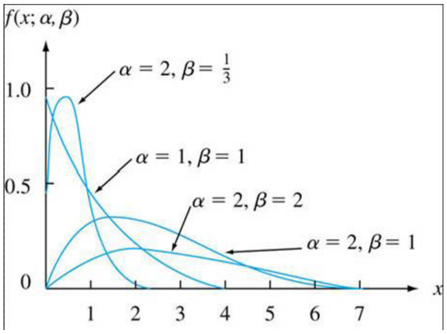
Chi-Squared Distribution: $X \sim \chi^2_\nu$

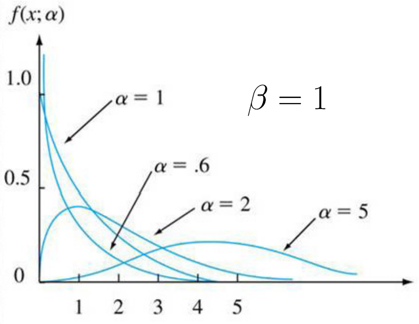
Section 6.3

Gamma Distribution

Def: A continuous RV X is said to have a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if the pdf of X is

$$f_X(x|\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

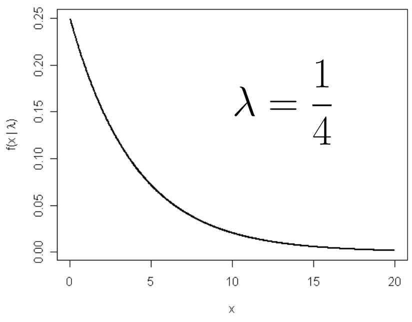


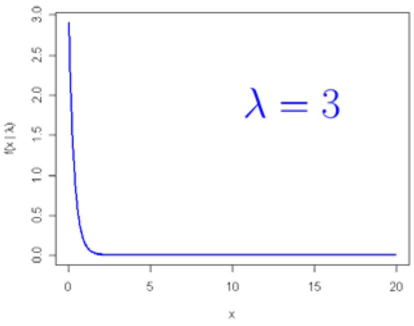


Exponential Distribution

Def: A continuous RV *X* follows an exponential distribution with parameter λ > 0 if *X* has pdf

$$f_X(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$





Chi-Squared Distribution

The Chi-Squared distribution : A special case of the Gamma distribution

when α = ν / 2 (for some positive integer ν) and β = 2.

The parameter ν is called the "degrees of freedom".

$$X \sim \chi^2_\nu \qquad f(x;\nu) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

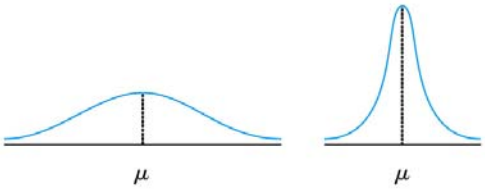
- Used for statistical inference on sampling distribution of Sample Variance of observations from a Normal Population

The Normal (Gaussian) Distribution

Notation: $X \sim \mathrm{N}(\mu, \sigma^2)$ ~ short hand for (is distributed as)

Probability density function (pdf):

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



© 2007 Thomson Higher Education

Section 6.5

17

The Standard Normal Distribution

Special case: μ = 0 and σ² = 1 (Standardized Scores)

$$\phi(z) \equiv f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$$Z \sim \mathrm{N}(0, 1)$$

cdf: $\Phi(z) \equiv \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw$

$$= \int_{-\infty}^{z} \phi(w)\, dw$$

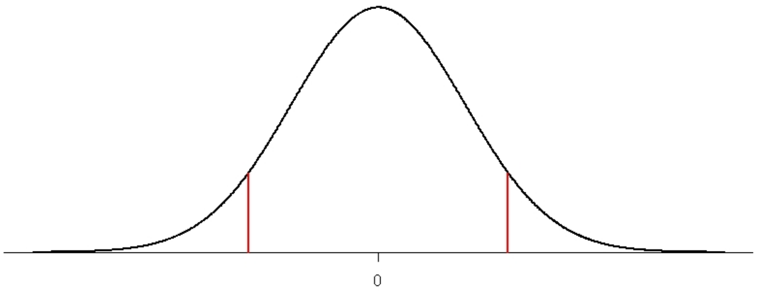
- Can not be expressed in closed functional form!
- USE Tables or Numerical Integration to evaluate this Integral

18

Standard Normal Distribution CDF

Table III provides values of area under the curve from 0 to z, for *z* = 0 (.01) 3.09, and z= 4.0, 5.0, 6.0

For, negative values of z, use the symmetry of the standard normal pdf.

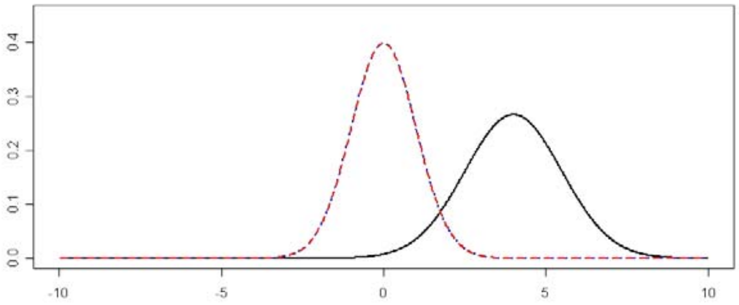


Transforming (Standardizing) Normal RV's

Idea: Transform a N(μ, σ^2) RV into a N(0, 1) RV...

$X \sim \mathrm{N}(\mu, \sigma^2)$ then: $W \equiv X - \mu \sim \mathrm{N}(0, \sigma^2)$

$$Z \equiv \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1)$$

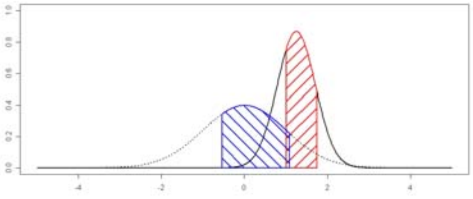


Using the Transformation

Say $X \sim \mathrm{N}(\mu, \sigma^2)$ and we want to compute $P(a \leq X \leq b)$

Idea: Transform to the standard normal distribution:

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$



21

Sampling Distributions

- Select a random sample of n observations from a specified population
  - Independent & identically distributed (i.i.d.) random vars
- These arise in a variety of experimental situations.
- Sampling distribution of a statistic:
  - Given repeated samples, the value of the statistic (e.g. sample mean) varies from sample to sample. The Sampling Distribution describes the pattern of variability in the values over all possible samples of a fixed size.

Chapter 8, Section 8.1

Sampling Distributions - 2

- If we know the underlying population distribution f, we can obtain the EXACT SAMPLING distribution of the sum (or, average)
- Sometimes, it is not easy to find this distribution analytically. Approximations via
  - Monte Carlo simulations
  - Large sample limiting distribution

Sampling Distribution of the Sample Mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We will focus on the probability distribution of $\overline{X}_n$ in two situations:

1. $X_i \sim N(\mu, \sigma^2)$ “No approximation needed.”

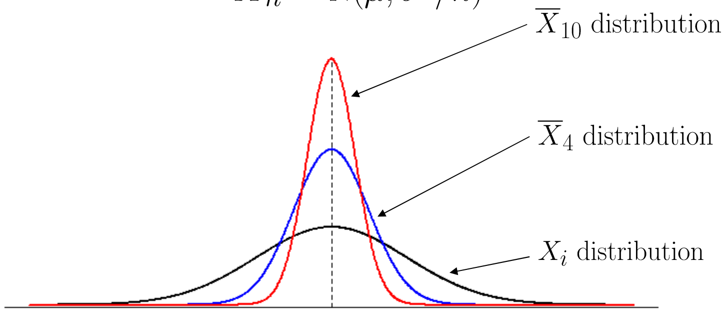
2. X_i’s have an arbitrary (but same) underlying probability distribution For large sample size, use “Limit distribution ”: $n \to \infty$

- In statistical inference problems, f is usually unknown. For large n, we are able to approximate this sampling distribution without knowing f
  Modes of convergence of Sample Mean

Situation 1: Sample from a Normal population

$$X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2), \quad i = 1, \ldots, n$$

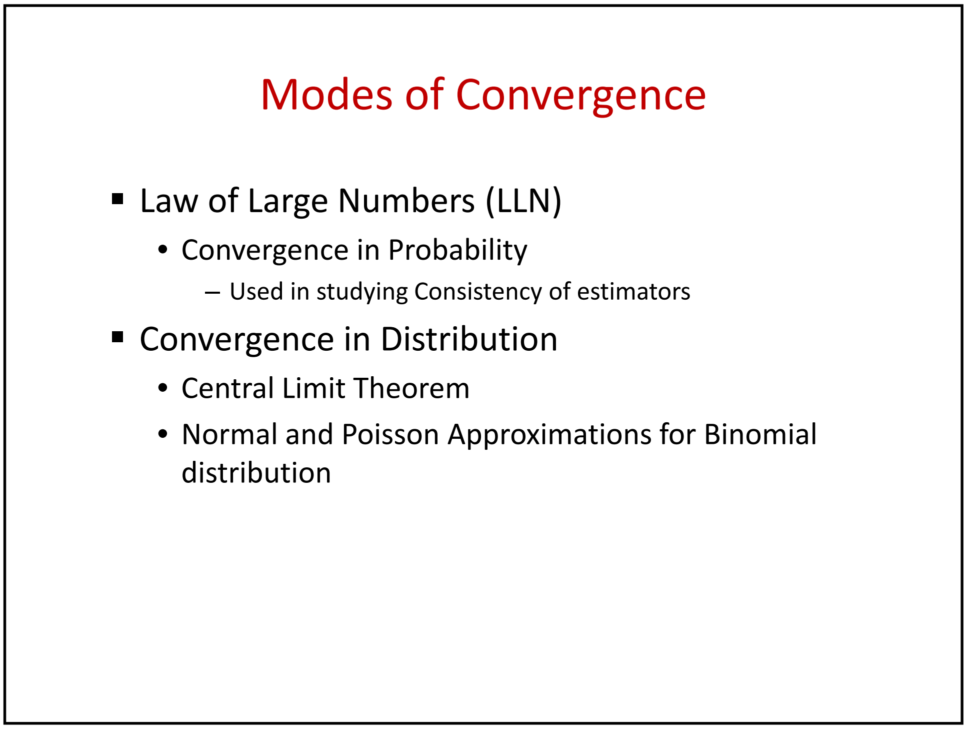
In this case: $\overline{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$



25

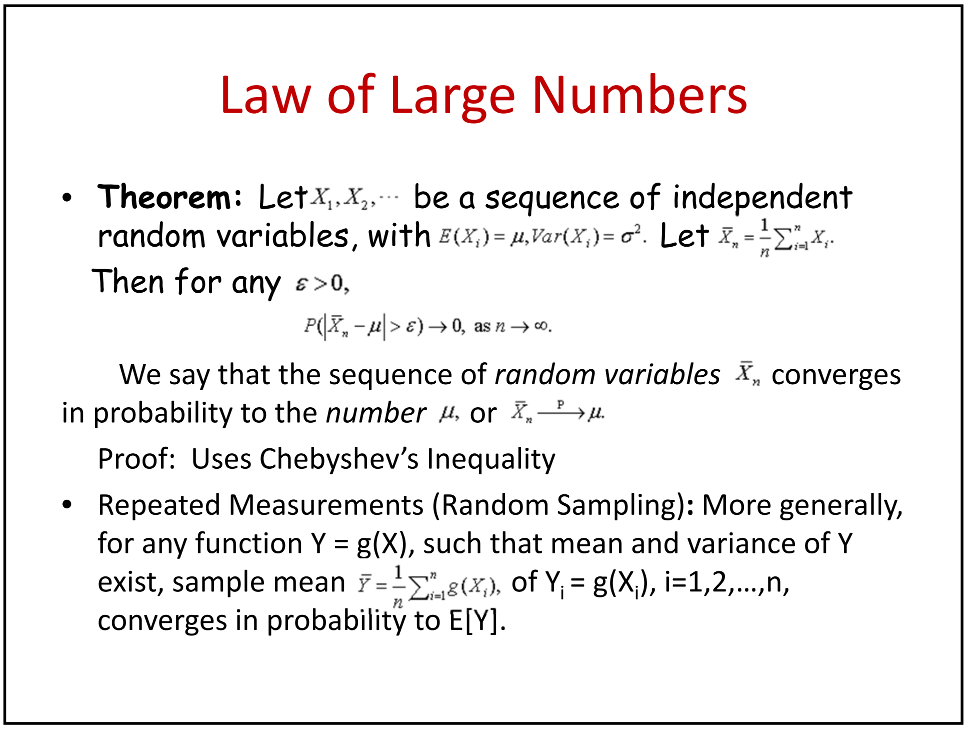
Convergence for Large Sample Size

- **Example:** Toss a coin a large number of, say n, times. How does one formalize the phrase *Proportion of heads is ~ ½*?
- Suppose that $X_1, X_2, \cdots$ is a sequence of independent Bernoulli trials, each with probability of success p. Then E(X_i) = p.
- The proportion of successes in n trials = $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- This sequence is random, in that over repeated trials, the sequence will takes different values.
- However, as n gets large, it does converge in *some* well defined sense.



Modes of Convergence

- Law of Large Numbers (LLN)
  - Convergence in Probability
    - Used in studying Consistency of estimators
- Convergence in Distribution
  - Central Limit Theorem
  - Normal and Poisson Approximations for Binomial distribution



Law of Large Numbers

- **Theorem:** Let $X_1, X_2, \cdots$ be a sequence of independent random variables, with $E(X_i) = \mu, Var(X_i) = \sigma^2$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

  Then for any $\varepsilon > 0$,

$$P(|\bar{X}_n - \mu| > \varepsilon) \to 0, \text{ as } n \to \infty.$$

We say that the sequence of *random variables* $\bar{X}_n$ converges in probability to the *number* μ, or $\bar{X}_n \xrightarrow{P} \mu$.

Proof: Uses Chebyshev's Inequality

- Repeated Measurements (Random Sampling)**:** More generally, for any function Y = g(X), such that mean and variance of Y exist, sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$, of $Y_i = g(X_i)$, i=1,2,…,n, converges in probability to E[Y].

Why do we need to look beyond LLN?

- LLN implies that one can estimate population mean and second moment quite well from a sample of n independent draws, if n is reasonably large.
- The chance that error in sample mean beyond any specified threshold get smaller and smaller as n increases.
- However, it does not allow us to assess the size of error, when one uses sample average to estimate the population mean.
  - If we want to approximate probability of a given size of error, we need to zoom into these errors on a more and more micro-scale c_n that goes to zero as n gets large.
  - Thus we consider a normalized quantity $U_n = \frac{(\bar{X}_n - \mu)}{c_n}$, whose distribution, f_n, does not degenerate to zero.

Convergence of means of a random sample of n observations: Central Limit Theorem

Second Case:

$X_1, \ldots, X_n$ are iid with mean μ and variance σ^2.

- We want to understand how the error { $\bar{X}_n - \mu$ } fluctuates around zero over repeated sampling.
- When *n* is sufficiently large, we say

  $\overline{X}_n$ is approximately distributed as N(μ,σ^2/n).

- In what sense? Consider the standardized random variable

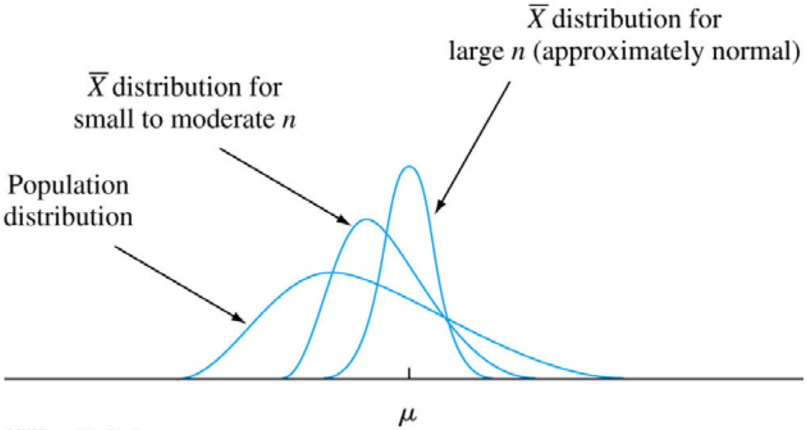
$$Z_n = \frac{(\bar{X}_n - \mu)}{(\sigma / \sqrt{n})} = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

  - Note that E[Z_n] = 0, Var[Z_n] = 1.

Insight on CLT

- The CLT asserts that the cdf of Z_n converges to $\Phi(\cdot)$, the cdf of a standard normal random variable Z for all real numbers, so we can approximate
- $P(Z_n \leq z) \doteq \Phi(z).$
- The proof in the book is under more restrictive conditions assuming that X_i's have same distribution, and its mgf exists. But this is not necessary. CLT holds under very general conditions.

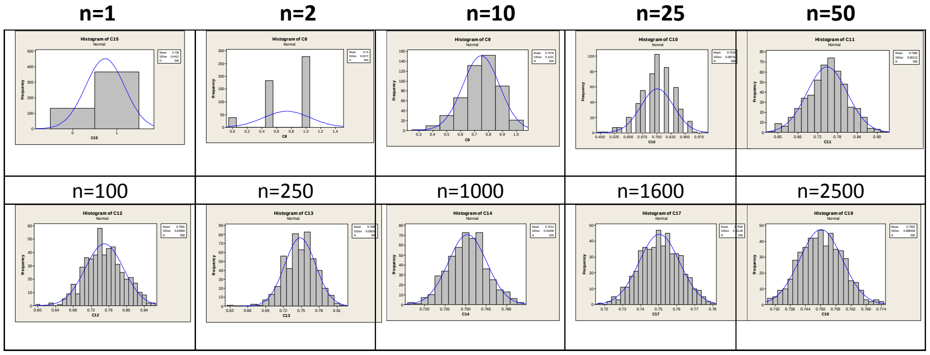
Illustrations



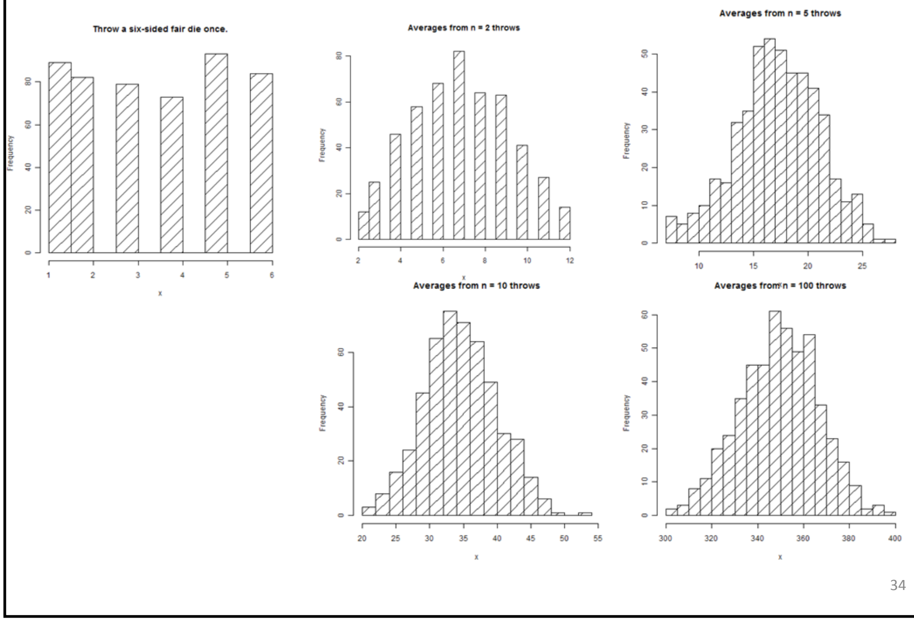
© 2007 Thomson Higher Education

Summary Statistics for 500 Repeated Samples of Averages of Various Sample Sizes

- **Underlying Population: Bernoulli (p=.75)**

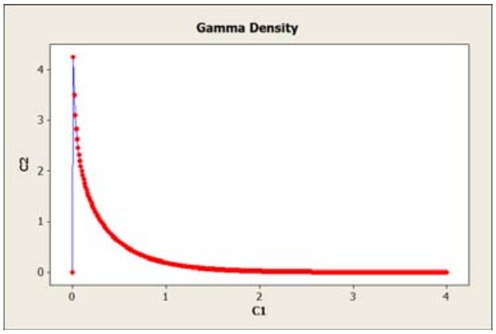


Sum of n Throws for a six sided fair die

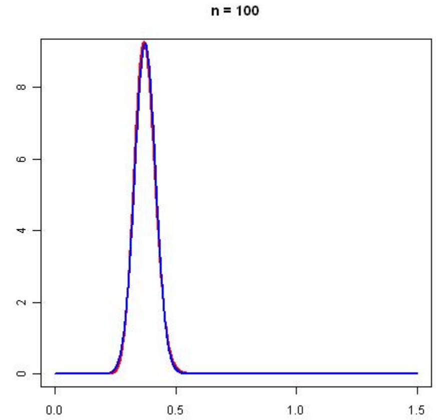


Sample Means from Gamma Distributed Random Variable

- X ~ Gamma ($\alpha = 0.75$, $\beta = 0.5$)
- The population pdf is sketched below:



Sampling distribution of the average

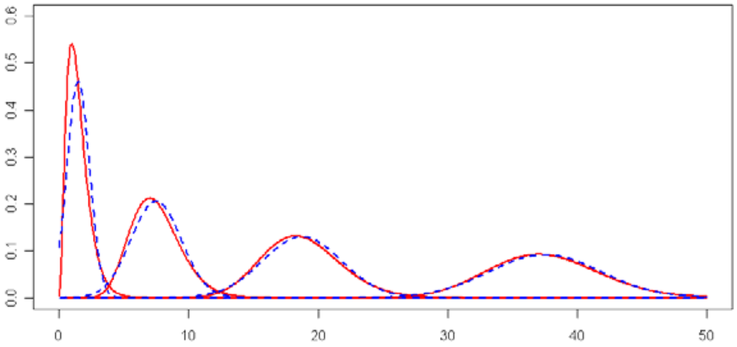


Central Limit Theorem for Totals

$X_1, \ldots, X_n$ are iid with mean μ and variance σ^2.

Define the sum of these *n* random variables: $T_n = \sum_{i=1}^{n} X_i$

When *n* is sufficiently large, $T_n \overset{\text{approx.}}{\sim} \mathrm{N}(n\mu, n\sigma^2)$


0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
50