Week 9
The Central Limit Theorem and Estimation Concepts
Week 9 Objectives

1. The Law of Large Numbers and the concept of consistency of averages are introduced. The condition of existence of the population mean is elaborated.
   - The Cauchy distribution is used to illustrate the effects of violating this condition.

2. The averages of $n = 2$ and of $n = 3$ uniform random variables are used to demonstrate that the exact distribution of averages depends on the sample size.

3. Simulations illustrate the approximate normality of averages.

4. The Central Limit Theorem, and the DeMoivre-Laplace Theorem are given.

5. Some estimation concepts are presented.
Consistency and Law of Large Numbers

The Central Limit Theorem
- Convolutions
- The Main Result
- The DeMoivre-Laplace Theorem

Lab 5: The distribution of averages

Some Estimation Concepts
- Model-Free and Model-Based Estimators
- Comparing Estimators
- The Method of Moments
We have seen that $\overline{X}$, $\hat{\rho}$ and $S^2$, approximate their population counterparts and that the approximation improves as the sample size increases. This is called **consistency**.

In general, an estimator $\hat{\theta}$ of a population parameter $\theta$ is called **consistent** if

$$
\hat{\theta} \to \theta \quad \text{as } n \text{ increases to } \infty.
$$

Consistency is such a basic property that all estimators used in statistics are consistent.

Thus, the sample covariance, sample correlation, and the estimators of the intercept and slope of the simple linear regression model are all consistent estimators.
Consistency is supported by a very famous probability result:

**Theorem (The Law of Large Numbers)**

Let $X_1, \ldots, X_n$ be iid and let $g$ be a function such that $-\infty < E[g(X_1)] < \infty$. Then,

$$\hat{\theta} \equiv \frac{1}{n} \sum_{i=1}^{n} g(X_i) \to \theta \equiv E[g(X_1)] \text{ as } n \text{ increases to } \infty.$$  

- Consistency of the aforementioned estimators follows from the *consistency of averages* stated in the LLN.
- For example, the consistency of $S^2$ follows from the fact that $\sigma^2 = E(X^2) - \mu^2$ and the consistency of $\bar{X}$ and $\frac{1}{n} \sum_{i=1}^{n} X_i^2$ as estimators of $\mu$ and $E(X^2)$, respectively.
The Assumption of finite $\mu$ and $\sigma^2$

- The LLN requires that the population has finite mean.
- Later we will be using also the assumption of finite variance.
- Is it possible for the mean of a r.v. to not exist or to be infinite? Or is it possible a r.v. to have finite $\mu$ but infinite $\sigma^2$?

The most notorious (for its abnormality) r.v. is the Cauchy whose pdf is

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}, \quad -\infty < x < \infty.$$ 

The median of this r.v. is zero, but its mean does not exist.
It is not hard to construct other PDFs with $\infty$ mean or variance:

Example

Let $X_1$ and $X_2$ have PDFs

$$f_1(x) = x^{-2}, \ 1 \leq x < \infty, \ f_2(x) = 2x^{-3}, \ 1 \leq x < \infty,$$

respectively. Show that $E(X_1) = \infty$, $E(X_2) = 2$, $\text{Var}(X_2) = \infty$.

Solution: By straightforward integration we get

$$E(X_1) = \int_1^\infty xf_1(x) \, dx = \infty, \ E(X_1^2) = \int_1^\infty x^2f_1(x) \, dx = \infty,$$

$$E(X_2) = \int_1^\infty xf_2(x) \, dx = 2, \ E(X_2^2) = \int_1^\infty x^2f_2(x) \, dx = \infty.$$

Thus, $\text{Var}(X_2) = E(X_2^2) - [E(X_2)]^2 = \infty$. 
Distributions with infinite mean, or infinite variance, are described as **heavy tailed**.

The LLN may not hold for heavy tailed distributions: \( \bar{X} \) may not converge (as in the case of Cauchy) or diverge. (Also the CLT, which we will see shortly, does not hold.)

Samples coming from heavy tailed distributions are much more likely to contain outliers.

If large outliers exist in a data set, it might be a good idea to focus on estimating another quantity, such as the median, which is well defined also for heavy tailed distributions.
Consistency and Law of Large Numbers

1. Consistency and Law of Large Numbers
2. The Central Limit Theorem
   - Convolutions
   - The Main Result
   - The DeMoivre-Laplace Theorem
3. Lab 5: The distribution of averages
4. Some Estimation Concepts
   - Model-Free and Model-Based Estimators
   - Comparing Estimators
   - The Method of Moments

Week 9 The Central Limit Theorem and Estimation Concepts
The LLN says nothing about the quality of the approximation for a given sample size. For that we need the distribution of \( \bar{X} \).

- The distribution of the sum of two independent random variables is called the *convolution* of the two distributions.
- In some cases it is easy to find the convolution of two RVs.

You and a friend each flip a fair coin a number of times. Let \( X \) be the number of heads you get in the \( n_1 \) flips, and \( Y \) be the number of heads your friend gets in \( n_2 \) flips.

(a) Are \( X \) and \( Y \) independent?
(b) What is the distribution of the total number of heads \( X + Y \)?

**Solution:** (a) Yes. (b) \( \sim X + Y \text{Bin}(n_1 + n_2, 0.5) \) (why?).

- In general, if \( X \sim \text{Bin}(n_1, p) \) and \( Y \sim \text{Bin}(n_2, p) \) are independent, then \( X + Y \sim \text{Bin}(n_1 + n_2, p) \).
The convolution is also known in the Poisson and Normal cases:

- If $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ are independent then $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. (See Example 5.3-1, p. 216 for the derivation.)
- If $X_1 \sim \text{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \text{N}(\mu_2, \sigma_2^2)$ are independent then $X_1 + X_2 \sim \text{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Such nice results are the exception. In general, the distribution of a sum of independent random variables is not easy to determine.

- If $X \sim \text{Bin}(n_1, p_1)$ and $Y \sim \text{Bin}(n_2, p_2)$, with $p_1 \neq p_2$, then the distribution of $X + Y$ is not binomial.
The convolution of two independent $U(0, 1)$ random variables is shown in
http://personal.psu.edu/acq/401/fig/convol_2Unif.pdf
This is derived in Example 5.3-3, p. 217. A numerical derivation will be done in the next lab.

In general, the formula for the convolution of two random variables is given in relation (5.3.3), p. 219. It is not easy to compute!

The convolution of several random variables is even harder to compute.
**Proposition**

Let $X_1, X_2, \ldots, X_n$ be independent with $X_i \sim N(\mu_i, \sigma_i^2)$, and set $Y = a_1 X_1 + \cdots + a_n X_n$. Then

$$Y \sim N(\mu_Y, \sigma_Y^2),$$

where

$$\mu_Y = a_1 \mu_1 + \cdots + a_n \mu_n,$$

and

$$\sigma_Y^2 = a_1^2 \sigma_1^2 + \cdots + a_n^2 \sigma_n^2.$$

**Corollary**

Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. Then

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$
Why care about such results?

- This result is relevant for answering an important statistical question, namely how to control the estimation error via the sample size:

  - We want to be 95% certain that $\overline{X}$ does not differ from the population mean by more than 0.3 units. How large should the sample size $n$ be? It is known that the population distribution is normal with $\sigma^2 = 9$.

Remark: This example assumes normality with known variance. In real life the variance is also unknown. More details on this will be given in Chapter 8.
Solution: Using the previous corollary we have

\[ P(\left| \bar{X} - \mu \right| < 0.3) = P \left( \frac{\left| \bar{X} - \mu \right|}{\sigma/\sqrt{n}} < \frac{0.3}{\sigma/\sqrt{n}} \right) = P \left( |Z| < \frac{0.3}{\sigma/\sqrt{n}} \right). \]

We want to find \( n \) so that \( P(\left| \bar{X} - \mu \right| < 0.3) = 0.95. \) Thus,

\[ \frac{0.3}{\sigma/\sqrt{n}} = z_{0.025}, \text{ or } n = \left( \frac{1.96}{0.3\sigma} \right)^2 = 384.16. \]

Thus, using \( n = 385 \) will satisfy the precision objective.
1 Consistency and Law of Large Numbers

2 The Central Limit Theorem
   - Convolutions
   - The Main Result
   - The DeMoivre-Laplace Theorem

3 Lab 5: The distribution of averages

4 Some Estimation Concepts
   - Model-Free and Model-Based Estimators
   - Comparing Estimators
   - The Method of Moments
The complexity of finding the exact distribution of a sum (or average) of $k \ (> 2)$ RVs increases rapidly with $k$.

This is why the following is the most important theorem in probability and statistics.

**Theorem (The Central Limit Theorem)**

Let $X_1, \ldots, X_n$ be iid with mean $\mu$ and variance $\sigma^2 < \infty$. Then, for large enough $n$ ($n \geq 30$ for the purposes of this course), $\overline{X}$ has approximately a normal distribution with mean $\mu$ and variance $\sigma^2/n$, i.e.

$$\overline{X} \sim N \left( \mu, \frac{\sigma^2}{n} \right).$$
The Central Limit Theorem: Comments

- The CLT does not require the $X_i$ to be continuous RVs.
- $T = X_1 + \cdots + X_n$ has approximately a normal distribution with mean $n\mu$ and variance $n\sigma^2$, i.e.

$$T = X_1 + \cdots + X_n \sim N\left(n\mu, n\sigma^2\right).$$

- The CLT can be stated for more general linear combinations of independent r.v.s, and also for certain dependent RVs. The present statement suffices for the range of applications of this book.
- The CLT explains the central role of the normal distribution in probability and statistics.
Example

Two towers are constructed, each by stacking 30 segments of concrete. The height (in inches) of a randomly selected segment is uniformly distributed in $(35.5, 36.5)$. A roadway can be laid across the two towers provided the heights of the two towers are within 4 inches. Find the probability that the roadway can be laid.

*Solution:* The heights of the two towers are

\[ T_1 = X_{1,1} + \cdots + X_{1,30}, \text{ and } T_2 = X_{2,1} + \cdots + X_{2,30}, \]

where $X_{i,j}$ is the height of the $j$th segment used in the $i$th tower. By the CLT,

\[ T_1 \sim N(1080, 2.5), \text{ and } T_2 \sim N(1080, 2.5). \]
Example (Continued)

Since the two heights $T_1$, $T_2$ are independent r.v.s, their difference, $D$, is distributed as

$$D = T_1 - T_2 \sim N(0, 5).$$

Therefore, the probability that the roadway can be laid is

$$P(-4 < D < 4) = P\left(\frac{-4}{2.236} < Z < \frac{4}{2.236}\right)$$

$$= \Phi(1.79) - \Phi(-1.79) = .9633 - .0367$$

$$= .9266.$$
Example

The waiting time for a bus, in minutes, is uniformly distributed in [0, 10]. In five months a person catches the bus 120 times. Find the 95th percentile, \( t_{0.05} \), of the person’s total waiting time \( T \).

**Solution.** Write \( T = X_1 + \ldots + X_{120} \), where \( X_i \) = waiting time for catching the bus the \( i \)th time. Since \( n = 120 \geq 30 \), and

\[
E(X_i) = 5, \quad \text{Var}(X_i) = \frac{10^2}{12} = \frac{100}{12},
\]

we have

\[
T \sim N \left( 120 \times 5, \ 120 \frac{100}{12} \right) = N(600, 1000).
\]

Therefore,

\[
t_{0.05} \approx 600 + z_{0.05} \sqrt{1000} = 600 + 1.645 \sqrt{1000} = 652.02.
\]
Consistency and Law of Large Numbers

The Central Limit Theorem

- Convolutions
- The Main Result
- The DeMoivre-Laplace Theorem

Lab 5: The distribution of averages

Some Estimation Concepts

- Model-Free and Model-Based Estimators
- Comparing Estimators
- The Method of Moments
The DeMoivre-Laplace theorem was proved first by DeMoivre in 1733 for $p = 0.5$ and extended to general $p$ by Laplace in 1812.

Due to the prevalence of the Bernoulli distribution in the experimental sciences, it continues to be stated separately even though it is now recognized as a special case of the CLT.

**Theorem (DeMoivre-Laplace)**

If $T \sim \text{Bin}(n, p)$ then, for $n$ satisfying $np \geq 5$ and $n(1 - p) \geq 5$, $T \sim N(np, np(1 - p))$. 
For an integer valued RV $X$, and $Y$ the approximating normal RV, the principle of continuity correction specifies

$$P(X \leq k) \approx P(Y \leq k + 0.5).$$

The DeMoivre-Laplace theorem and continuity correction yield

$$P(T \leq k) \approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right).$$

For individual probabilities use

$$P(X = k) \approx P(k - 0.5 < Y < k + 0.5).$$
Figure: Approximation of $P(X \leq 5)$ with and without continuity correction.
Example

A college basketball team plays 30 regular season games, 16 of which are against class A teams and 14 are against class B teams. The probability that the team will win a game is 0.4 if the team plays against a class A team and 0.6 if the team plays against a class B team. Assuming that the results of different games are independent, approximate the probability that

1. the team will win at least 18 games;
2. the number of wins against class B teams is smaller than that against class A teams.
Solution: If $X_1$ and $X_2$ denote the number of wins against class A teams and class B teams,

$$X_1 \sim \text{Bin}(16, 0.4), \quad \text{and} \quad X_2 \sim \text{Bin}(14, 0.6).$$

Since $16 \times 0.4$ and $14 \times 0.4$ are both $\geq 5$,

$$X_1 \sim N(16 \times 0.4, 16 \times 0.4 \times 0.6) = N(6.4, 3.84),$$

$$X_2 \sim N(14 \times 0.6, 14 \times 0.6 \times 0.4) = N(8.4, 3.36).$$

Consequently, since $X_1$ and $X_2$ are independent,

$$X_1 + X_2 \sim N(6.4 + 8.4, 3.84 + 3.36) = N(14.8, 7.20),$$

$$X_2 - X_1 \sim N(8.4 - 6.4, 3.84 + 3.36) = N(2, 7.20).$$
Solution continued: Hence, using also continuity correction, the needed approximation for part 1 is

\[ P(X_1 + X_2 \geq 18) = 1 - P(X_1 + X_2 \leq 17) \approx 1 - \Phi \left( \frac{17.5 - 14.8}{\sqrt{7.2}} \right) \]

\[ = 1 - \Phi(1.006) = 1 - 0.843 = 0.157, \]

and the needed approximation for part 2 is

\[ P(X_2 - X_1 < 0) \approx \Phi \left( \frac{-0.5 - 2}{\sqrt{7.2}} \right) = \Phi(-0.932) = 0.176. \]
Consistency/inconsistency results

- Consistency of the sample proportion
  
n=100; x=rbinom(1,n,0.3); x/n # Repeat with n=1000 and n=10000

- Consistency of the sample mean for normal samples
  
n=100; mean(rnorm(n)) # Repeat with n=1000 and n=10000

- Inconsistency of the sample mean when E(X) does not exist
  
n=100; mean(rcauchy(n)) # Repeat with n=1000 and n=10000
Plotting probabilities

- Probabilities can be plotted with `plot` and `barplot` commands. Empirical probabilities can be plotted with `hist` and `barplot`.

- Plot the pmf of the Bin(10, 0.5):
  ```r
  plot(0:10, dbinom(0:10, 10, 0.5), pch=4, col=2)
  lines(0:10, dbinom(0:10, 10, 0.5), col=2)
  ```
  Alternatively: 
  ```r
  p=dbinom(0:10, 10, 0.5); barplot(p)
  ```
  Try also:
  ```r
  barplot(p, xlim=c(0,12), ylim=c(0,0.25), col=2)
  ```

- Plot empirical probabilities resulting from 
  `x=rbinom(1000,10,0.5)`:
  ```r
  hist(x, seq(-0.5, 10.5, 1), col=5)
  ```
  Alternatively:
  ```r
  ph=table(x)/1000; barplot(ph)
  ```
The pdf of the sum of two independent uniform random variables can be obtained from the formula (5.3.3) in the book, i.e., $f_{X_1+X_2}(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x)f_{X_1}(x)dx$ and plotted by

$$g=\text{function}(x)\{x*(x<1) + (2-x)*(1<= x)*(x<2)\}$$

curve(g,0,2)

Formula (5.3.3) can be applied repeatedly to get the pdf of the sum of 3 or more uniform r.v.s, but it becomes very difficult quickly.

So, we employ simulations to get an idea what the pdf of the sum (or average) of several uniform r.v.s looks like:
m = matrix(runif(5000000), nrow=80)

hist(m[1,], seq(0, 1, 0.1)) # should look like the pdf of the uniform

hist((m[1,]+m[2,])/2, seq(0, 1, 0.01)) # should look like the plotted pdf

hist(colMeans(m), seq(min(colMeans(m))-.01, max(colMeans(m))+.01, 0.01)) # shows you what the density of \( \bar{X} \) for \( n = 80 \) looks like.

mean(m[1,]); sd(m[1,]) # should be approximately 1.6487 and \( \sqrt{6.389} = 2.5276 \)

mean(colMeans(m)); sd(colMeans(m)) # should be approximately 1.6487 and \( \sqrt{6.389/80} = 0.2826 \)
Averages of log-normal random variables

curve(dlnorm, 0,10) # This density has $\mu = \exp(0.5) = 1.6487$
and $\sigma^2 = \exp(2) - 1 = 6.389$

m=matrix(rlnorm(5000000), nrow=80)

hist(m[1,], seq(0, max(m[1,])+0.3, 0.3)) # should look like the plotted pdf

hist(colMeans(m), seq(min(colMeans(m))-0.1, max(colMeans(m))+0.1, 0.1)) # shows you what the density of $X$ for $n = 80$ looks like.

mean(m[1,]); sd(m[1,]) # should be approximately 1.6487 and $\sqrt{6.389} = 2.5276$

mean(colMeans(m)); sd(colMeans(m)) # should be approximately 1.6487 and $\sqrt{6.389}/80 = 0.2826$
Averages of negative binomial random variables

\[ p = \text{dnbinom}(0:15, 5, 0.6); \text{barplot}(p) \] # has
\[ \mu = 5(1/0.6) - 5 = 3.3333 \text{ and } \sigma^2 = 5(0.4/0.6^2) = 5.5556 \]

\[ m = \text{matrix} \left( \text{rnbinom}(4000000, 5, 0.6), \text{nrow}=80 \right) \]

\[ \text{hist}(m[1,], \text{seq}(-0.5, \text{floor}(\text{max}(m[1,]))+0.5,1)) \] # should look like the plotted pmf.

\[ \text{hist}(\text{colMeans}(m), \text{seq}(\text{floor}(\text{min}(\text{colMeans}(m)))-0.5, \text{floor}(\text{max}(\text{colMeans}(m)))+1.5, 0.1)) \] # shows you what the pmf of \( \bar{X} \) for \( n = 80 \) looks like.

\[ \text{mean}(m[1,]); \text{sd}(m[1,]) \] # should be approximately 3.3333 and \( \sqrt{5.5556} = 2.357 \)

\[ \text{mean}(\text{colMeans}(m)); \text{sd}(\text{colMeans}(m)) \] # should be approximately 3.3333 and \( \sqrt{5.5556/80} = 0.2635 \)
1. Consistency and Law of Large Numbers

2. The Central Limit Theorem
   - Convolutions
   - The Main Result
   - The DeMoivre-Laplace Theorem

3. Lab 5: The distribution of averages

4. Some Estimation Concepts
   - Model-Free and Model-Based Estimators
   - Comparing Estimators
   - The Method of Moments
In probability and statistics the word “model” refers to an assumption we make about the data at hand.

- For univariate discrete data we may assume the binomial, Poisson or another discrete probability model.
- For univariate continuous data we may assume the normal, exponential, or another continuous probability model.
- For bivariate data we may assume the bivariate normal model, the linear regression model, or another model.
When a model for the distribution of the data is assumed, it is of interest to estimate the model parameters.

- In the normal model we want to estimate $\mu$ and $\sigma^2$.
- In the uniform model, the two endpoints are of interest.
- In the SLR model we want $\beta_0, \beta_1$ and $\sigma^2_\epsilon$.

Estimating the model parameters is called fitting a model to data.

Estimating the model parameters is quite different from estimating the population mean, variance, percentiles etc we saw in Chapter 1.

- For example, how does one estimate the endpoints of a uniform distribution? (Chapter 6 presents two ways of doing this, and one of them will be discussed.)
Estimators of the model parameters are meaningful only if the assumed model is correct.

- Estimators of the endpoints of the uniform distribution would be meaningless if the data are not uniformly distributed.
- Estimators the slope and intercept of the SLR model lose their interpretation if the true regression function is a higher order polynomial.

The estimators of Chapter 1 do not have this problem.
- For example, assuming the population mean, $\mu$, exists, $\bar{X}$ estimates $\mu$ regardless of the population distribution.

The estimators of Chapter 1 will be called nonparametric or model-free.
Fitting models to data leads to alternative estimators of the population mean, variance, proportions and percentiles.

- If the normal model is assumed:
  a) The density is estimated by the $N(\bar{X}, S^2)$ density.
  b) The percentiles are also estimated as $\bar{X} + Sz_\alpha$.
  c) In addition to sample proportions, $P(X \leq x)$ may also be estimated by $\Phi((x - \bar{X})/S)$.

- If the uniform model is assumed, and $\hat{a}, \hat{b}$ are the estimators of the two endpoints:
  a) The mean and median are also estimated as $(\hat{a} + \hat{b})/2$.
  b) The variance is also estimated as $(\hat{b} - \hat{a})^2/12$.

- If the Poisson model is assumed, $\bar{X}$ is an alternative estimator of the variance.

Such alternative estimators are called **model-based**.
Consistency and Law of Large Numbers

The Central Limit Theorem
  - Convolutions
  - The Main Result
  - The DeMoivre-Laplace Theorem

Lab 5: The distribution of averages

Some Estimation Concepts
  - Model-Free and Model-Based Estimators
  - Comparing Estimators
  - The Method of Moments
The existence of alternative estimators for the same population quantity raises the question of which is “better”.

- Estimators are compared in terms of their mean and variance. [Piece of jargon: The standard deviation of an estimator is called its standard error.]
- If the mean of an estimator $\hat{\theta}$ equals its true value, i.e.,
  \[ E_{\theta} \left( \hat{\theta} \right) = \theta, \]
  it is called unbiased.
- The difference expected value $-$ true value is the bias:
  \[ \text{bias}(\hat{\theta}) = E_{\theta} \left( \hat{\theta} \right) - \theta \]
$\overline{X}$ and $\hat{p}$ are unbiased estimators:

$$E_{\mu}(\overline{X}) = \mu, \quad E_{\rho}(\hat{p}) = p.$$

$S^2$ is also unbiased: $E_{\sigma^2}(S^2) = \sigma^2$. See p. 232 for a proof.

Later we will also see that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased.

However,

- $S$ is only consistent (not unbiased) estimator of $\sigma$.
- $\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2$ is only consistent (not unbiased) estimator of $\sigma^2$. 
The Mean Square Error Selection Criterion

- Variance and bias combine to define the Mean Square Error, $\text{MSE}_{\theta}(\hat{\theta}) = E_{\theta}(\hat{\theta} - \theta)^2$:

  $$\text{MSE}_{\theta}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2$$

- The MSE selection criterion says that estimator $\hat{\theta}_1$ should be preferred over $\hat{\theta}_2$ if

  $$\text{MSE}_{\theta}(\hat{\theta}_1) \leq \text{MSE}_{\theta}(\hat{\theta}_2)$$

- If $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased, we have the variance selection criterion: $\hat{\theta}_1$ should be preferred over $\hat{\theta}_2$ if

  $$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2).$$
On the basis of the MSE (or variance) selection criterion, we can make the following statement:

- Under correct model assumptions, model-based estimators are preferable to the model-free ones. For example:
  - if the normal assumption is correct, $\overline{X} + Sz_\alpha$ is preferable to the $(1 - \alpha)100\%$ sample percentile.
  - If the normal assumption is correct, $\Phi((x - \overline{X})/S)$ is preferable to the sample proportion.
  - If the normal assumption is correct, $\overline{X}$ is preferable to $\widetilde{X}$.
  - If the Poisson assumption is correct, $\overline{X}$ is preferable (as estimator of the variance!) to $S^2$. 
Consistency and Law of Large Numbers

The Central Limit Theorem

1. Convolutions
2. The Main Result
3. The DeMoivre-Laplace Theorem

Lab 5: The distribution of averages

Some Estimation Concepts

1. Model-Free and Model-Based Estimators
2. Comparing Estimators
3. The Method of Moments
For $r \geq 1$, the $r$-th moment of a r.v. $X$ is defined as $E(X^r)$.

Let $X_1, \ldots, X_n$ be iid from the same population as $X$. Then,

- By the LLN, $\frac{1}{n} \sum_{i=1}^{n} X_i^r$ is a consistent estimator of $E(X^r)$.
- It is also unbiased.

The method of moments is based on expressing the model parameters in terms of the moments.

- If $X \sim \text{Unif}(0, \theta)$ then $\theta = 2\mu$.
- If $X \sim \text{Unif}(\alpha, \beta)$ then

  \[ \alpha = \mu - \sqrt{3\sigma^2}, \quad \beta = \mu + \sqrt{3\sigma^2}, \]

  and, of course, $\sigma^2 = E(X^2) - \mu^2$. 
Additional examples include:

- If $X \sim \text{Exp}(\lambda)$, $\lambda = 1/\mu$; if $X \sim \text{Poisson}(\lambda)$, $\lambda = \mu$.

Ignoring the difference between $\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2$ and $S^2$, the above lead to the following method of moments estimators:

- If $X_1, \ldots, X_n$ is a sample from $N(\mu, \sigma^2)$, then
  \[ \hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = S^2. \]

- If $X_1, \ldots, X_n$ is a sample from $\text{Unif}(0, \theta)$, $\hat{\theta} = 2\overline{X}$.
- If $X_1, \ldots, X_n$ is a sample from $\text{Unif}(\alpha, \beta)$,
  \[ \hat{\alpha} = \overline{X} - \sqrt{3}S^2, \quad \hat{\beta} = \overline{X} + \sqrt{3}S^2, \]
- If $X_1, \ldots, X_n$ is a sample from $\text{Exp}(\lambda)$, $\hat{\lambda} = 1/\overline{X}$. 


In expressing the model parameters in terms of moments, the method of moments uses the lowest possible moments.

- If \( X_i, i = 1, \ldots, n \), are iid Poisson(\( \lambda \)), use \( \hat{\lambda} = \bar{X} \) instead of \( \hat{\lambda} = S^2 \).

- For bivariate data, \((X_i, Y_i), i = 1, \ldots, n\), the **product moment** \( E(XY) \), is estimated by \( \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \).

- The method of moments estimator of \( \text{Cov}(X, Y) \) is

\[
\frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \bar{X} \bar{Y},
\]

which is slightly different from the sample version of \( \text{Cov}(X, Y) \).