ABE6933 PSC, Fall 2020
Probability & Mathematical Statistics: a Scientific Computing Approach

Part III: Point Estimation, Limit Theorems

Instructor: Dr. Nikolay Bliznyuk

September 2, 2020

Parametric family of distributions

Notation: $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ is the parametric family of distributions indexed by θ.

Parameterization is the correspondence between θ and F_θ.

Common setup: $\underbrace{Y_1, Y_2, \ldots, Y_n}_{\text{iid sample from } F_\theta} \overset{\text{iid}}{\sim} F_\theta$.

WMS: $Y_1, Y_2, \ldots, Y_n$ is a "random sample" from F_θ.

Warning: in general, "random" does not mean "independent".

Goals of probability and statistics

Goal of probability: determine $\Pr(T(Y_1, \ldots, Y_n) \in A)$, where T is some function.

To motivate **goals of statistics**, consider a game:

1. Mother Nature chooses $\theta \in \Theta$ and generates $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F_\theta$.
2. Goal of statistics/statistician: use info in the sample $Y_1, \ldots, Y_n$ to make inference/learn/guess the value of θ that the nature has chosen.

What is meant by statistical inference?

Inference = estimation + hypothesis testing

Estimation:

1. Point estimation: use of data $Y_1, Y_2, \ldots, Y_n$ to "guess" θ using a rv $T(Y_1, \ldots, Y_n)$ that is close to θ in some prob. sense.
2. Set/interval estimation: find a random set/interval $S(Y_1, \ldots, Y_n)$ such that $\Pr(\theta \in S(Y_1, \ldots, Y_n))$ is high.

Hypothesis testing: use the sample to determine if a hypothesis $\theta = \theta_0$ is likely to be true. "Do the data support the hypothesis that $\theta = \theta_0$?"

Other goals of statistics (besides inference): modeling, prediction.

Check out the video about Ritz Casino scam for an amazing illustration of how modeling and prediction are useful in real life `https://www.youtube.com/watch?v=GnaOM4W-hDE`

Identifiability of a parameterization

Recall the **goal of statistics** and the game:

1. Mother Nature chooses $\theta \in \Theta$ and generates $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F_\theta$.
2. Goal of statistics/statistician: use info in the sample $Y_1, \ldots, Y_n$ to make inference/learn/guess the value of θ that the nature has chosen, or the distribution F_θ.

A parameterization $\theta \mapsto F_\theta$ is called *identifiable* if $\theta \neq \theta'$ implies $F_\theta \neq F_{\theta'}$.

Point Estimation: Master Plan

1. Basic characteristics of point estimators (bias, MSE, efficiency). Finding unbiased estimators.

2. Estimators of “popular” functions.

3. Methods of finding estimators: method of moments (MOM) and maximum likelihood estimation (MLE).

4. Large-sample (aka asymptotic) properties of estimators (consistency, asymptotic normality). Convergence in probability and in distribution.

What is a statistic?

Def. A *statistic* T is an observable function of the random sample, i.e., $T = T(Y_1, \ldots, Y_n)$, where the function T does not depend on unknown parameters.

Def. $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$ is the sample mean.

Def. $S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$ is the sample variance (notice the square to distinguish it from the sum S_n).

Examples

Estimators and estimates

Def. An *estimator* of $g(\theta)$ is a statistic $T(Y_1, Y_2, \ldots, Y_n)$ that is used to "guess" the value of $g(\theta)$.

Def. An *estimate* of $g(\theta)$ is the value of the estimator $T(Y_1, \ldots, Y_n)$ when $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.

Notice: $T(Y_1, \ldots, Y_n)$ is a rv (an estimator), while $T(y_1, \ldots, y_n)$ is a fixed number (an estimate).

Our interest is in finding good estimators.

Two principal methods are the method of moments and the maximum likelihood estimation (discussed later).

To compare the "goodness" estimators, it is necessary to define several criteria, below. It is assumed here that θ is a scalar; this can be generalized to the case when θ is a vector.

Bias

Def. *Bias* of an estimator $\widehat{\theta}$ of θ is

$$Bias(\theta|\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Notice that, typically, the bias is a function of θ (and, possibly, of other parameters).

If $Bias(\theta|\widehat{\theta}) = 0$ for every value of θ, the estimator $\widehat{\theta}$ is called *unbiased*.

Examples:

Finding unbiased estimators

Suppose θ is a scalar and $\widehat{\theta}_n$ is some estimator of θ such that $\mathrm{E}\left(\widehat{\theta}_n\right) = a + b\theta$, where a and b do not depend on θ and $b \neq 0$. Then $\mathrm{E}\left((\widehat{\theta}_n - a)/b\right) = \theta$ and hence $(\widehat{\theta}_n - a)/b$ is unbiased for θ.

Example: Let $Y_1, \ldots, Y_n \sim$ Uniform$(0, \theta)$. Find unbiased estimators of θ based on $\bar{Y}_n$ and $Y_{(n)} = \max(Y_1, \ldots, Y_n)$.

MSE and its bias-variance decomposition

Def. *Mean squared error* of an estimator $\widehat{\theta}$ is

$$MSE(\theta|\widehat{\theta}) = E\{(\widehat{\theta} - \theta)^2\}.$$

Bias-variance decomposition of the MSE:

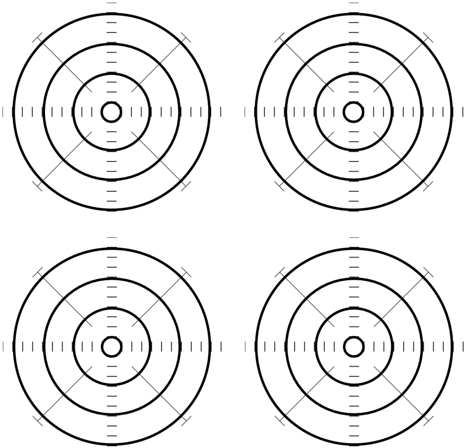
$$MSE(\theta|\widehat{\theta}) = \{Bias(\theta|\widehat{\theta})\}^2 + Var(\widehat{\theta}).$$

Consequence: bias-variance tradeoff.

Precision and accuracy

Precision is the reciprocal of the variance.

An estimator is called *precise* if its variance is low. An estimator is called *accurate* if its MSE (i.e., both variance and bias) is low.



MLE: Intuition via Bayes Rule I

$$\begin{aligned}
Pr(Z = z|Y = y) &= \frac{Pr(Z = z, Y = y)}{Pr(Y = y)} \\
&= \frac{Pr(Y = y, Z = z)}{\sum_{t \in \mathcal{Z}} Pr(Y = y, Z = t)} \\
&= \frac{Pr(Y = y|Z = z)Pr(Z = z)}{\sum_{t \in \mathcal{Z}} Pr(Y = y|Z = t)Pr(Z = t)}.
\end{aligned}$$

Here, $Pr(Z = z)$ is the *prior* probability of the event $\{Z = z\}$, while $Pr(Z = z|Y = y)$ is the *posterior* probability of the event $\{Z = z\}$ given the "experimental evidence" $\{Y = y\}$.

E.g., flip a coin $n = 100$ times independently with the probability of success Z; observe $\{Y = y\}$ successes (e.g., $y = 67$).

Q: What is your best guess about the true probability of success Z, given that you observed y successes?

MLE: Intuition via Bayes Rule II

E.g., flip a coin $n = 100$ times independently with the probability of success Z; observe $Y = y$ successes (e.g., $y = 67$). Suppose Z is a rv such that $Pr(Z = i/100) = 1/101$ for $i = 0, 1, \ldots, 100$.

Q: What is the most likely value of Z, given that we observed y successes?

Principle of Maximum Likelihood Estimation: discrete rvs

Let $y_1, y_2, \ldots, y_n$ be the observed values of iid rvs $Y_1, Y_2, \ldots, Y_n$.

When the Y_i's are discrete rvs,

$$Pr(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} Pr(Y_i = y_i) = \prod_{i=1}^{n} f(y_i|\theta) > 0$$

is the probability of observing the vector of outcomes $[y_1, y_2, \ldots, y_n]$.

Since the events of high probability are more likely to occur than the events of low probability, and the event $[Y_1 = y_1, \ldots, Y_n = y_n]$ has occurred, it is sensible to estimate the unknown parameter θ using the value $\widehat{\theta}(y_1, \ldots, y_n)$ that makes $P(Y_1 = y_1, \ldots, Y_n = y_n|\theta)$ as high as possible.

Q: What if Y_i's are indep., have different distr's?
A: Replace f(y_i|theta) by f_i(y_i|theta).

Principle of ML Estimation: continuous rvs

Let $y_1, y_2, \ldots, y_n$ be the observed values of iid rvs $Y_1, Y_2, \ldots, Y_n$.

When the Y_i's are continuous rvs, $Pr(\bigcap_{i=1}^n [Y_i = y_i] | \theta) = 0$. However, in this case

$$\prod_{i=1}^n f(y_i|\theta) \approx \frac{Pr(\bigcap_{i=1}^n [y_i - \delta/2 \leq Y_i \leq y_i + \delta/2] \, |\theta)}{\delta^n} > 0.$$

Hence maximization of $\prod_{i=1}^n f(y_i|\theta)$ wrt θ is equivalent to maximization wrt θ the probability of the event that $\bigcap_{i=1}^n \{Y_i \in [y_i - \delta/2, y_i + \delta/2]\}$.

Method of ML Estimation: Preliminaries

Likelihood function is the joint probability density or mass function of the data, treated as a function of θ, i.e.,

$$L(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{i}(y_i|\theta).$$

Notice that, in $L(\theta|y_1, \ldots, y_n)$, θ is the variable, and the sample $[y_1, \ldots, y_n]$ is treated as fixed.

Recall that in the pdf/pmf, θ is held fixed, and the y_i's vary. For convenience, $L(\theta|y_1, \ldots, y_n)$ will be abbreviated as $L(\theta)$.

Log-likelihood function is $l(\theta) = \log L(\theta)$. Here, $\log$ is typically taken to be the natural logarithm, $\ln$.

Example: Write down the likelihood and log-likelihood functions when $Y_1, \ldots, Y_n$ are iid $Bernoulli(p)$ rvs.

MLE: Procedure

Step 1: Write down the likelihood as a function of the parameter (vector) θ.

Step 2: Write down the log-likelihood as a function of the parameter (vector) θ, call it $l(\theta)$.

Step 3: Maximize the log-likelihood function with respect to θ. Often, but not always, this amounts to

Step 3a: solving for θ the *score equation*

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = 0.$$

Step 3b: if $\widehat{\theta}$ is the solution, checking that $\widehat{\theta}$ is indeed the maximizer of $l(\theta)$. Often, this amounts to checking that

$$\left. \frac{\partial^2 l(\theta)}{\partial \theta^2} \right|_{\theta=\widehat{\theta}} < 0.$$

Consistency and convergence in probability

An estimator $\widehat{\theta}_n$ is *consistent* for θ if and only if for every fixed tolerance $\delta > 0$

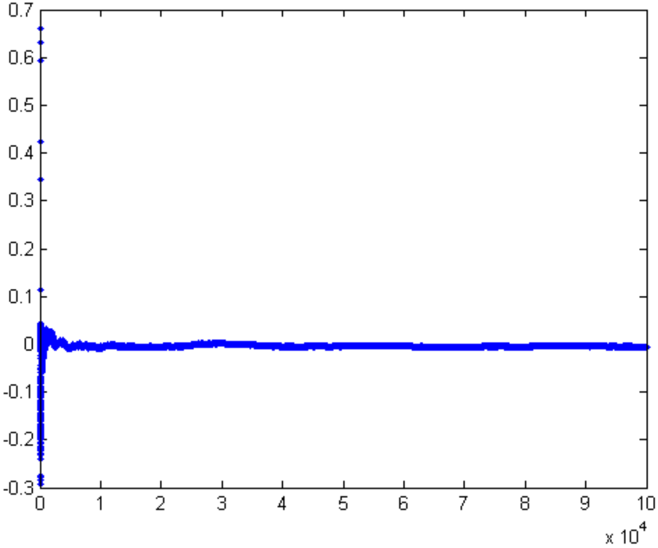
$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \delta) = 1 - \lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| \leq \delta) = 0.$$

An estimator $\widehat{\theta}_n$ is said to *converge in probability to θ*, denoted as

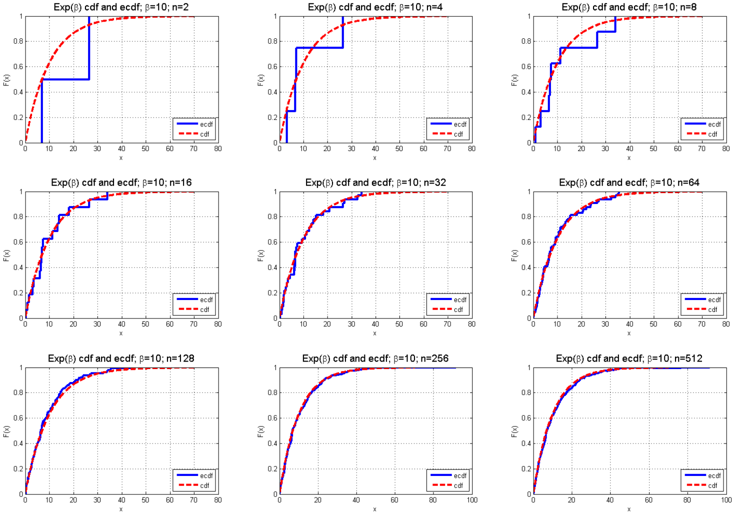
$$\widehat{\theta}_n \to^P \theta,$$

if and only if $\widehat{\theta}_n$ is consistent for θ.

Sample mean $\bar{Y}_n$ for a sequence of iid Normal(0,1) rvs



$$\widehat{F}_n(x) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(Y_i \leq x) \to^{P} F(x);\ F = Exp(\beta = 10)$$



Example: Laws of Large Numbers

Weak Law of Large Numbers (WLLN): If $Y_1, \dots, Y_n$ are iid with $E(Y_i^2) < \infty$, then $\bar{Y}_n \to^P E(Y_i)$.

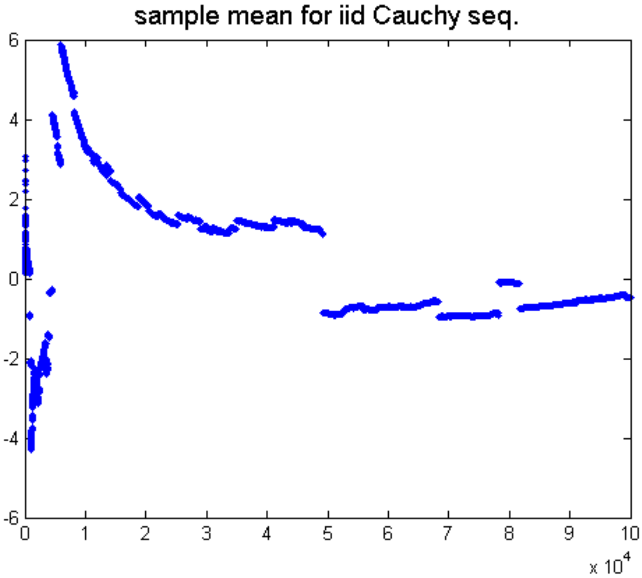
Strong Law of Large Numbers (SLLN): If $Y_1, \dots, Y_n$ are iid with $E(|Y_i|) < \infty$, then $\bar{Y}_n \to^P E(Y_i)$.

E.g.: Inconsistency of the sample mean of iid Cauchy rvs

if $Y_1, \ldots, Y_n$ are iid from the Cauchy distribution with density $f(x) = 1/\{\pi(1 + (x - \mu)^2)\}$, $\bar{Y}_n$ does not converge in probability to a constant. Here, μ happens to be the population median.

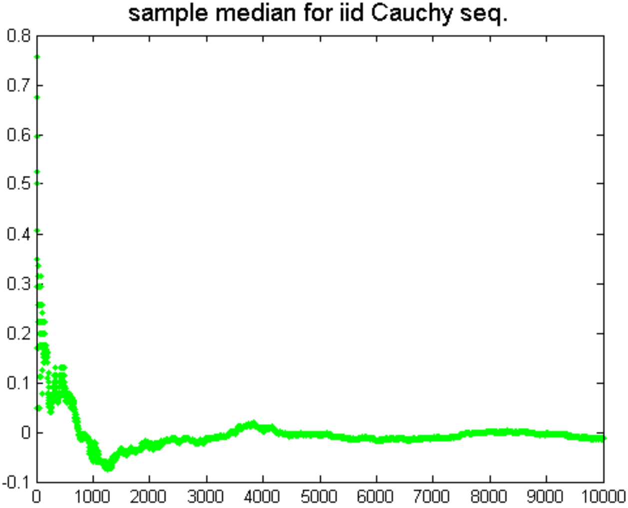
One can show analytically that $\bar{Y}_n$ has the same distribution as Y_1.

Sample means for a sequence of iid Cauchy rvs; true $\mu = 0$



E.g.: Consistency of the sample median with Cauchy rvs

If $Y_1, \ldots, Y_n$ are iid from the Cauchy distribution with $\mu = 0$, the sample median converges in probability to 0.



Convergence in Distribution

Let F_i be the cdf of a rv Y_i and let F_Y be the cdf of a rv Y.

A sequence of rvs $Y_1, Y_2, \ldots$ is said to *converge in distribution* to a rv Y, denoted as $Y_n \to^D Y$, if $F_i(x) \to F_Y(x)$ for every point x where F_Y is continuous.

In this case, F_Y is known as an *asymptotic* or *limiting* distribution of Y_n (or of the sequence $Y_1, Y_2, \ldots$).

Practical interpretation: for n sufficiently large, the cdf F_n of Y_n is "close" to F_Y.

Henceforth, if we do not know the cdf of Y_n but $Y_n \to^D$, then we may use F_Y to approximate the probabilities of the events of the form $[Y_n \leq x]$.

Central Limit Theorem (CLT)

Let $Y_1, Y_2, \ldots, Y_n$ are iid with $\mathrm{E}\left(Y_i^2\right) < \infty$, $\mathrm{E}\left(Y_i\right) = \mu$, $\mathrm{Var}\left(Y_i\right) = \sigma^2$. Then

$$Z_n = \frac{\overline{Y}_n - \mathrm{E}\left(\overline{Y}_n\right)}{\sqrt{\mathrm{Var}\left(\overline{Y}_n\right)}} = \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to a rv $Z \sim$ Normal$(0, 1)$.

Important: μ and σ^2 are not allowed to depend on n.

Practical Implications:

1. For large n, Z_n is approximately Normal$(0, 1)$.
2. For large n, $\overline{Y}_n$ is approximately distributed as
3. For large n, $S_n = \sum_{i=1}^{n} Y_i$ is approximately distributed as

More on the Assumptions of the CLT

1. $Y_1, \ldots, Y_n$ are iid with $\mathrm{E}(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$, i.e., no particular form of distribution is assumed. But if the Y_i's come from a parametric family, then μ and σ^2 will be functions of model parameters.

2. It is implicit that μ and σ^2 do not depend on n.

Significance of the CLT: for large n, $Z_n \overset{\mathrm{D}}{\approx}$ Normal$(0, 1)$.

The CLT ensures that Z_n, S_n and $\bar{Y}_n$ are approximately normal, regardless of the actual distribution of the Y_i's (only need to know the mean and variance).

Note: This assumes that n is "sufficiently large", which (unfortunately), does depend on the distribution of the Y_i's.

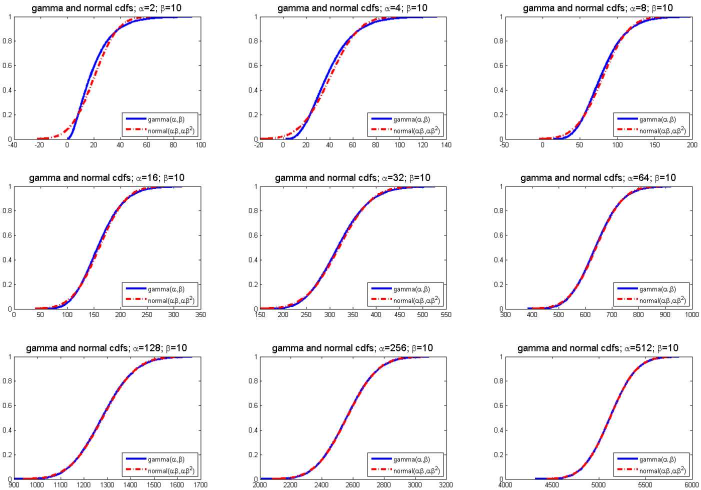
Illustration of the CLT

Let's find the exact distribution of $S_n = \sum_{i=1}^n Y_i$ (when this is possible) and compare it to the approximate distribution suggested by the CLT.

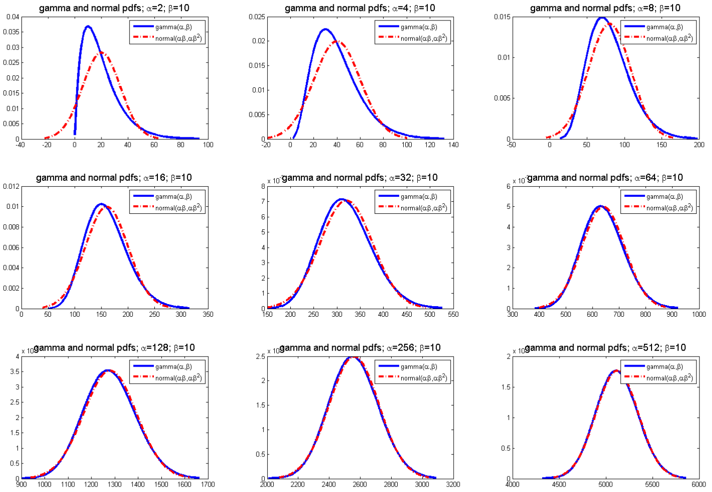
1. $Y_1, \ldots, Y_n \sim \text{Gamma}(\alpha, \beta)$. mgf of $Y_i : (1 - \beta t)^{-\alpha} = m(t)$.

2. $Y_1, \ldots, Y_n \sim \text{Poisson}(\lambda)$. mgf: $m(t) = \exp\big(\lambda(e^t - 1)\big)$.

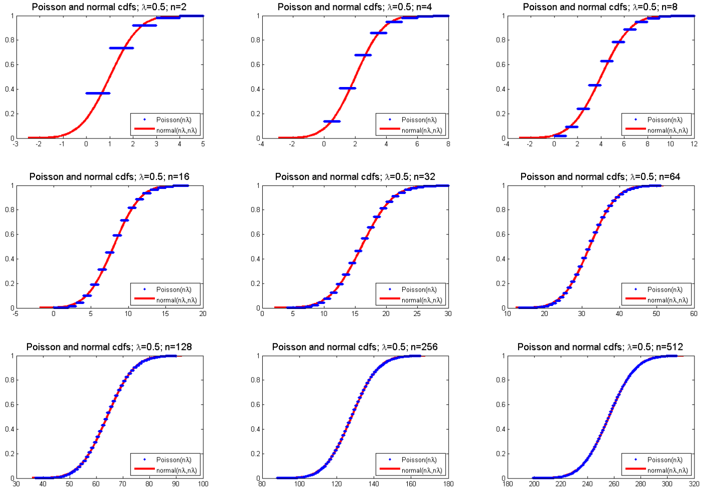
Exact and approx. cdfs of S_n when $Y_i \sim^{iid} Gamma(1, 10)$



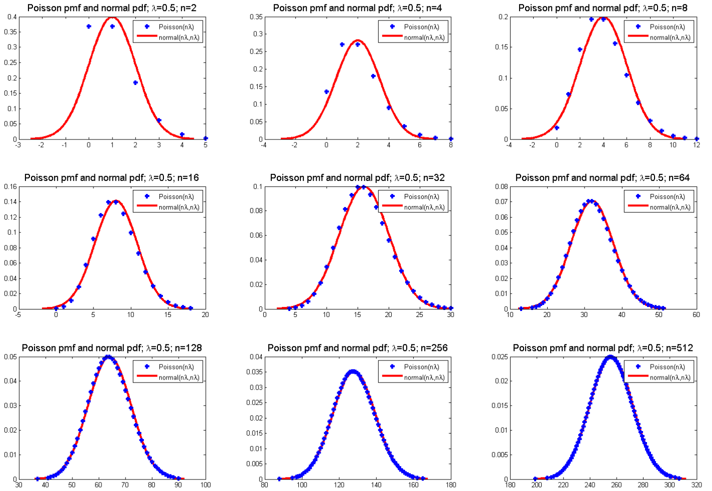
Exact and approx. pdfs of S_n when $Y_i \sim^{iid} Gamma(1, 10)$



Exact and approx. cdfs of S_n when $Y_i \sim^{iid} Poisson(\lambda)$



Exact and approx. pmfs of S_n when $Y_i \sim^{iid} Poisson(\lambda)$



When the CLT breaks down

CLT can break down when assumptions are not satisfied:

E.g. 1: Let $Y_1, \ldots, Y_n$ be iid Cauchy rvs. One can show that for any constants $a > 0, b > 0$, $aY_1 + bY_2 \stackrel{\mathrm{D}}{=} (a+b)Y_1$.

E.g. 2: Let $Y_1, \ldots, Y_n$ be iid $Bernoulli(p_n)$ where $p_n n \to \lambda$.
Then $S_n = \sum_{i=1}^n Y_i \to^D Poisson(\lambda)$.