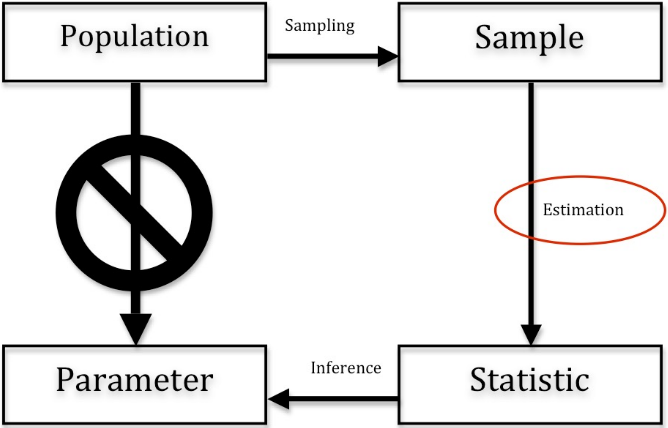
PLSC 502 – Fall 2024
Estimation and Estimators

October 21, 2024

Remember This?



Random Variables, Take Two

For a random variable X:

$$X_i = \underbrace{\mu}_{\text{“systematic part”}} + \underbrace{u_i}_{\text{“stochastic part”}}$$

where μ is the population mean (expected value) of X and $\text{Cov}(\mu, u) = 0$.

That implies that:

$$\underbrace{u_i}_{\text{“error”}} = \underbrace{X_i}_{\text{“observed”}} - \underbrace{\mu}_{\text{“expected”}}$$

Random Variables, Take Two

What's our expectation for u?

$$\begin{aligned}
\mathsf{E}(u) &= \mathsf{E}(X - \mu) \\
&= \mathsf{E}(X) - \mathsf{E}(\mu) \\
&= \mathsf{E}(X) - \mu \\
&= \mu - \mu \\
&= 0
\end{aligned}$$

and so:

$$\begin{aligned}
\mathsf{Var}(X) &= \mathsf{E}[(X - \mu)^2] \\
&= \mathsf{E}(u^2)
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{Var}(u) &= \mathsf{E}[(u - E(u))^2] \\
&= \mathsf{E}[(u - 0)^2] \\
&= \mathsf{E}(u^2).
\end{aligned}$$

Estimation Example: $\bar{X}$

Challenge: Estimate $\mu = E(X)$ from a sample of N observations.

$$
\begin{aligned}
\bar{X} &= \frac{1}{N}\sum_{i=1}^{N} X_i \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mu + u_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mu) + \frac{1}{N}\sum_{i=1}^{N}(u_i) \\
&= \frac{1}{N}(N\mu) + \frac{1}{N}\sum_{i=1}^{N}(u_i) \\
&= \mu + \bar{u}
\end{aligned}
$$

The point: $\bar{X}$ *is a random variable*.

Small-Sample Properties

- Hold irrespective of N
- "Small sample estimators"

Large-Sample (Asymptotic) Properties

- Hold as $N \to \infty$
- "More is better"

Unbiasedness

Start with a generic population parameter θ, and an estimator of it $\hat{\theta}$ based on a sample of N observations...

<u>Unbiasedness</u> means:

$$\mathrm{E}(\hat{\theta}) = \theta$$

"Bias" is:

$$B(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$$

Example: For $\bar{X}$, we know that:

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}(\mu + \bar{u}) \\
&= \mathrm{E}(\mu) + \mathrm{E}(\bar{u}) \\
&= \mu + 0 \\
&= \mu
\end{aligned}$$

and so:

$$B(\bar{X}) = 0.$$

Multiple Unbiased Estimators

For $N = 2$:

$$Z = \lambda_1 X_1 + \lambda_2 X_2.$$

note that

$$\begin{aligned}
E(Z) &= E(\lambda_1 X_1 + \lambda_2 X_2) \\
&= E(\lambda_1 X_1) + E(\lambda_2 X_2) \\
&= \lambda_1 E(X_1) + \lambda_2 E(X_2) \\
&= \lambda_1 \mu + \lambda_2 \mu \\
&= (\lambda_1 + \lambda_2)\mu
\end{aligned}$$

Means

$$E(Z) = \mu \;\longleftrightarrow\; (\lambda_1 + \lambda_2) = 1.0$$

and in fact:

$$E(Z) = \mu \;\longleftrightarrow\; \sum_{i=1}^{N} \lambda_i = 1.0.$$

Q: Why do we use $\lambda_i = \frac{1}{N}\ \forall\ i$?

Efficiency:

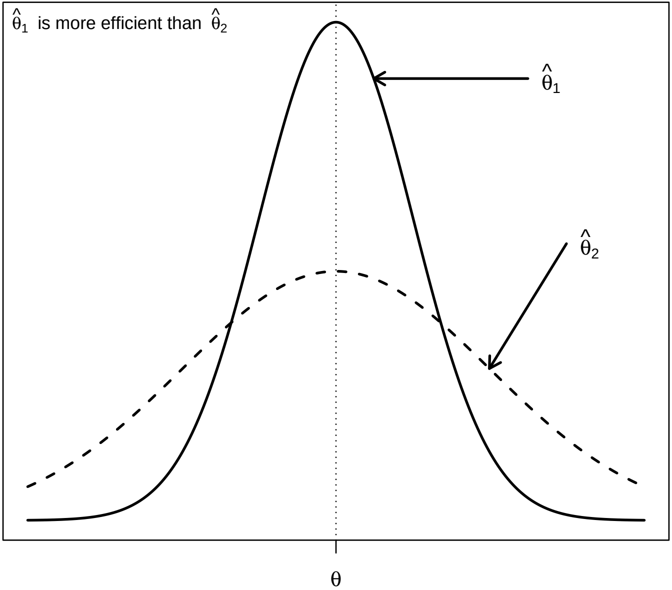
- is *relative variability* – how much difference we would expect in our $\hat{\theta}$s from one sample to the next...
- ...so a more efficient estimator has higher "reliability."
- ...is related to **information** (specifically, the *Fisher information* in the sample).

Note that:

- To be *fully efficient*[1], an estimator must be unbiased. BUT...
- ...the least-variance estimator need not be an unbiased one.

[1] That is, to achieve the *Cramer-Rao lower bound*, something we'll discuss in detail a bit later.

Efficiency: Unbiased $\hat{\theta}$s



Efficiency (continued)

Note that for our example with $N = 2$, where $\text{Var}(X) = \sigma^2$:

$$\begin{aligned}\text{Var}(Z) &= \text{Var}(\lambda_1 X_1 + \lambda_2 X_2) \\ &= (\lambda_1^2 + \lambda_2^2)\sigma^2\end{aligned}$$

and:

$$\begin{aligned}\lambda_1^2 + \lambda_2^2 &= \lambda_1^2 + (1 - \lambda_1)^2 \\ &= \lambda_1^2 + (1 - 2\lambda_1 + \lambda_1^2) \\ &= 2\lambda_1^2 - 2\lambda_1 + 1.\end{aligned}$$

Minimize!

$$\frac{\partial 2\lambda_1^2 - 2\lambda_1 + 1}{\partial \lambda_1} = 4\lambda_1 - 2$$

$$\begin{aligned}4\lambda_1 - 2 &= 0 \\ \lambda_1 &= 0.5\end{aligned}$$

Mean Squared Error

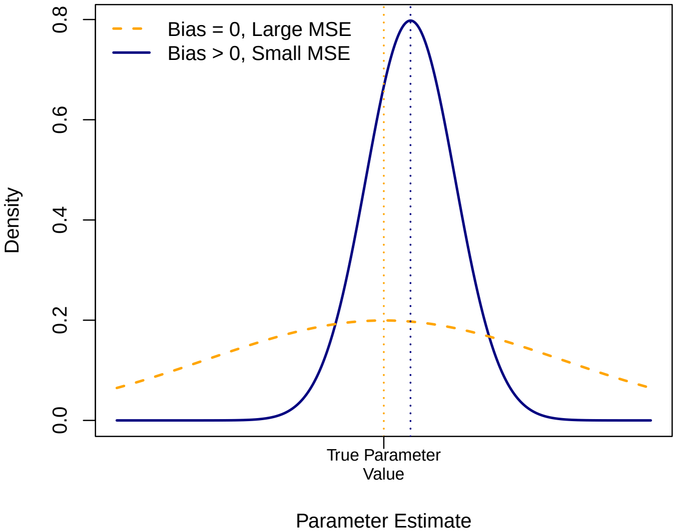
The "mean squared error" ("MSE") of an estimator $\hat{\theta}$ is:

$$\begin{aligned} \mathsf{MSE}(\hat{\theta}) &= \mathsf{E}[(\hat{\theta} - \theta)^2] \\ &= \mathsf{E}[B(\hat{\theta})^2] \\ &= \mathsf{Var}(\hat{\theta}) + [B(\hat{\theta})^2] \end{aligned}$$

Note that:

- The MSE of an unbiased estimator is equal to its variance [that is, MSE $=$ Var$(\hat{\theta})$].
- Among unbiased estimators, the efficient estimator will always have the smallest MSE [because $B(\hat{\theta}) = [B(\hat{\theta})]^2 = 0$].

MSE Illustrated



Comparing Estimators via MSE

As an estimator of μ, $\bar{X}$ has:

- $B(\bar{X}) = 0$
- $\text{Var}(\bar{X}) = \sigma^2/N$, so
- $\text{MSE}(\bar{X}) = \sigma^2/N + (0)^2 \ = \ \sigma^2/N$.

My alternative: the "Six Estimator"!

$$\zeta = 6$$

(That's a "zeta." Gotta learn your Greek letters.)

Comparing Estimators via MSE

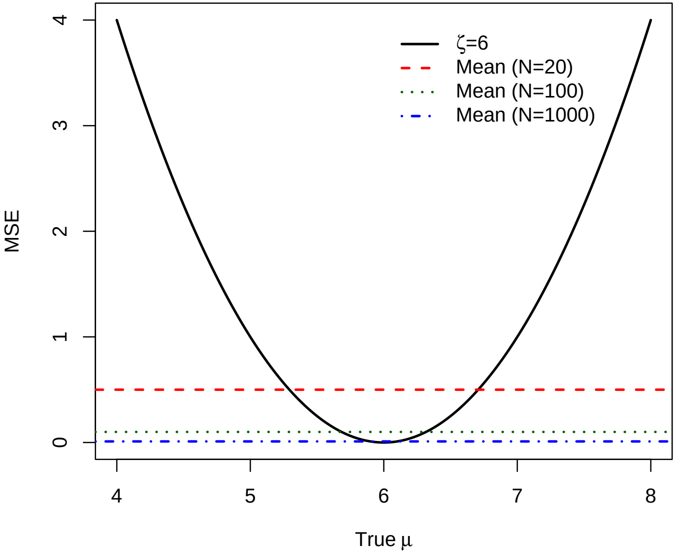
Properties of ζ (for $\zeta = 6$):

$$\begin{aligned} B(\zeta) &= \mathsf{E}(\zeta - \mu) \\ &= \mathsf{E}(6) - \mathsf{E}(\mu) \\ &= 6 - \mu, \end{aligned}$$

$$\begin{aligned} \mathsf{Var}(\zeta) &= \mathsf{Var}(6) \\ &= 0 \end{aligned}$$

and so:

$$\begin{aligned} \mathsf{MSE}(\zeta) &= \mathsf{Var}(\zeta) + [B(\zeta)]^2 \\ &= 0 + (6 - \mu)^2 \\ &= 36 - 12\mu + \mu^2 \end{aligned}$$



The black line is the MSE of ζ, expressed as a function of the "true" population mean μ. The other colored lines are the MSEs for $\bar{X}$, under the assumption that $\sigma^2 = 10$ and $N = \{20, 100, 1000\}$, respectively.

Large-Sample Properties: Consistency

An estimator $\hat{\theta}$ is *consistent* if:

$$\lim_{N \to \infty} \text{Pr}[|\hat{\theta} - \theta| < \epsilon] = 1.0$$

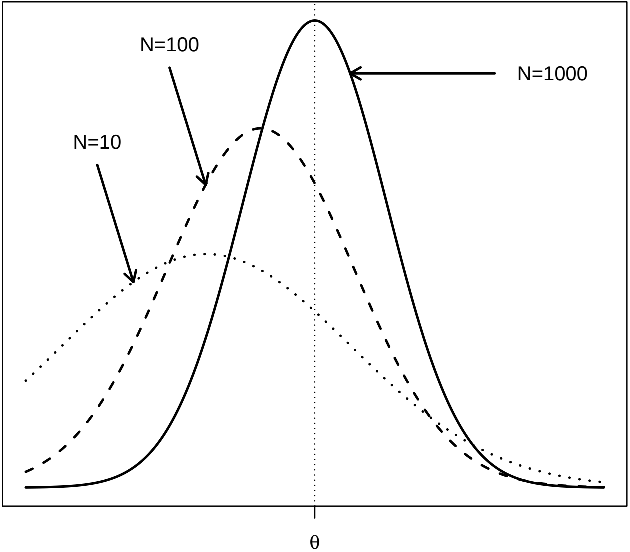
for an arbitrarily small $\epsilon > 0$

Equivalently:

$$\text{E}(\hat{\theta}_N) \to \theta \text{ as } N \to \infty$$

Intuition: "Asymptotic unbiasedness"...

A Consistent Estimator $\hat{\theta}$



Estimation, Generally

Among estimators:

- Unbiased $>$ Consistent $>$ Biased
- Fully Efficient $>$ Asymptotically Efficient $>$ Inefficient
- MSE is <u>one</u> way to trade off bias vs. efficiency

Estimation Example: The Poisson

Recall the *Poisson* distribution:

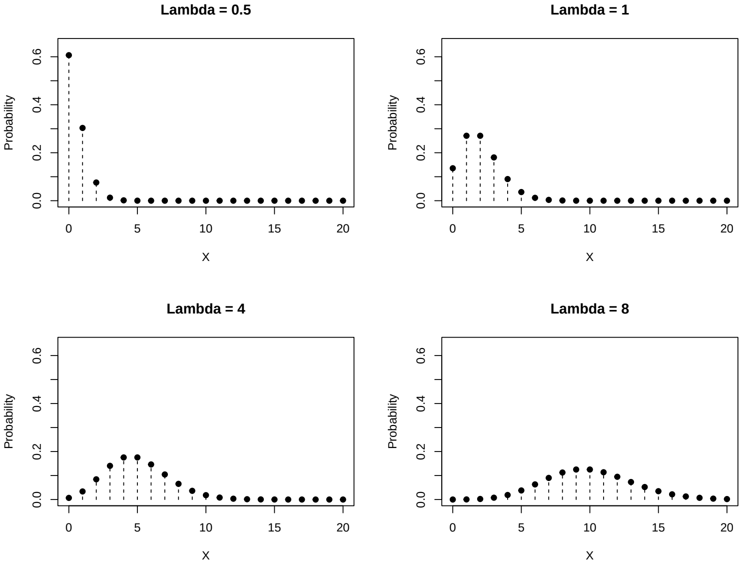
$$f(x) \equiv \Pr(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}.$$

for $x \in \{0, 1, 2, ...\}$.

The Poisson:

- ...is a distribution for *counts* of *independent events*;
- ...is a *one parameter* distribution, where
- ...the parameter λ is both the *mean* and the *variance* of X.

Poisson Densities



Poisson Estimation

<u>What is a "good" estimator for λ?</u>

For a series of N i.i.d. values $\{X_1, X_2, ...X_N\}$ drawn from a Poisson distribution, their *joint* probability is:

$$f(X_1, X_2, ...X_N|\lambda) \equiv f(\mathbf{X}) = \prod_{i=1}^{N} \frac{\lambda^{X_i} \exp(-\lambda)}{X_i!}. \qquad (1)$$

This is sometimes known as the *likelihood* (more on that later...), and it relies on the fact that the joint probability of two independent random variables equals the product of the two marginal probabilities:

$$\Pr(A, B \mid A \perp B) = \Pr(A) \times \Pr(B)$$

We can simplify (1) by taking its log:

$$\begin{aligned}
\ln[f(\mathbf{X})] &= \ln\left[\prod_{i=1}^{N} \frac{\lambda^{X_i}\exp(-\lambda)}{X_i!}\right] \\
&= \sum_{i=1}^{N} \ln\left[\frac{\lambda^{X_i}\exp(-\lambda)}{X_i!}\right] \\
&= \sum_{i=1}^{N} [X_i \ln(\lambda) - \lambda - \ln(X_i!)] \\
&= -N\lambda + \ln(\lambda)\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \ln(X_i!)
\end{aligned}$$

(This is the *log-likelihood*...)

Poisson Estimation

If we want to know the value of λ that maximizes this joint (log-)probability, we can figure that out too:

$$\frac{\partial \ln f(\mathbf{X})}{\partial \lambda} = -N + \frac{1}{\lambda}\sum_{i=1}^{N} X_i$$

and then:

$$-N + \frac{1}{\lambda}\sum_{i=1}^{N} X_i = 0$$

and so:

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

IOW, one version of a "good" estimator for λ (the "maximum likelihood estimator") is the empirical mean $\bar{X}$...

Poisson Mean Characteristics

What can we say about this $\hat{\lambda}$?

$$
\begin{aligned}
E(\hat{\lambda}) &= E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} E(X_i) \\
&= \frac{1}{N}\sum_{i=1}^{N} \lambda \\
&= \lambda
\end{aligned}
$$

so:

$$B(\hat{\lambda}) = 0 \ \text{(unbiasedness)}$$

Also: Because $\mathrm{Var}(X) = \lambda$, this also means that $\hat{\lambda}$ is also an unbiased estimate of the variance.

More Poisson Mean Characteristics

Variance / efficiency?

Because $\hat{\lambda}$ is unbiased, we know that:

$$\mathsf{MSE}(\hat{\lambda}) = \mathsf{Var}(\hat{\lambda}).$$

Central limit theorem means that:

$$\hat{\lambda} \sim N\left(\lambda, \frac{\lambda}{N}\right)$$

so:

$$\mathsf{MSE}(\hat{\lambda}) = \frac{\lambda}{N}.$$

Example One: Simulation

The Plan:

1. Draw N values of X from a Poisson distribution with a known value of λ;
2. Calculate $\hat{\lambda} = \bar{X}$;
3. Repeat steps (1) - (2) many times;
4. Examine the distribution of the $\hat{\lambda}$s

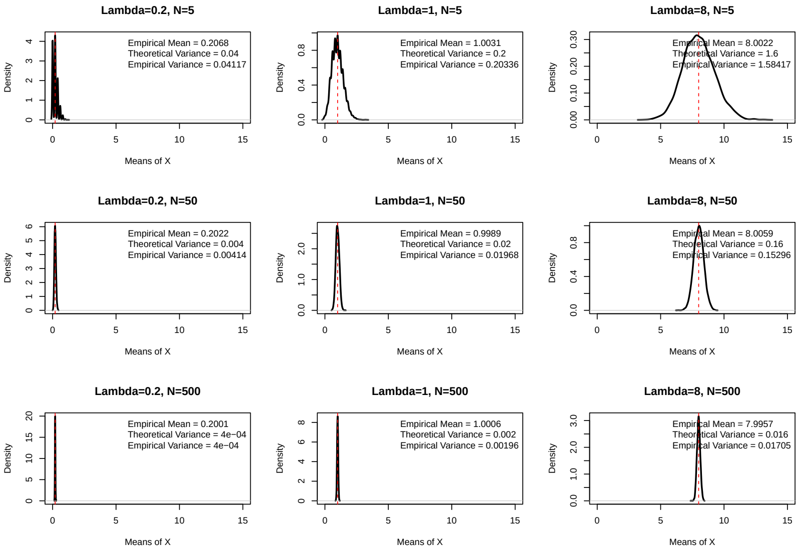
Details

- Vary $\lambda \in \{0.2, 1.0, 8.0\}$
- Vary $N \in \{5, 50, 500\}$

A Little Code

```
> L<-c(0.2,1,8)  # the lambdas:
> N<-c(5,50,500) # the Ns:
> sims<-4000     # number of sims
> Out<-data.frame(matrix(nrow=sims,ncol=length(N)*length(L)))
>
> c <- 0                         # column indicator for "Out"
> set.seed(7222009)              # Seed
>
> for(i in 1:length(N)) {        # Looping over sample sizes...
+   for(j in 1:length(L)) {      # Looping over lambdas
+     c <- c+1                   # increment column indicator
+     for(k in 1:sims) {         # Looping over 4000 simulations each
+       df<-rpois(N[i],L[j])     # Draw N values from Poisson(lambda)
+       Out[k,c]<-mean(df)       # Store the mean of the N draws
+       rm(df)
+     }
+   }
+ }
```

A Picture


Lambda=0.2, N=5
Empirical Mean = 0.2068
Theoretical Variance = 0.04
Empirical Variance = 0.04117
Lambda=1, N=5
Empirical Mean = 1.0031
Theoretical Variance = 0.2
Empirical Variance = 0.20336
Lambda=8, N=5
Empirical Mean = 8.0022
Theoretical Variance = 1.6
Empirical Variance = 1.58417
Lambda=0.2, N=50
Empirical Mean = 0.2022
Theoretical Variance = 0.004
Empirical Variance = 0.00414
Lambda=1, N=50
Empirical Mean = 0.9989
Theoretical Variance = 0.02
Empirical Variance = 0.01968
Lambda=8, N=50
Empirical Mean = 8.0059
Theoretical Variance = 0.16
Empirical Variance = 0.15296
Lambda=0.2, N=500
Empirical Mean = 0.2001
Theoretical Variance = 4e−04
Empirical Variance = 4e−04
Lambda=1, N=500
Empirical Mean = 1.0006
Theoretical Variance = 0.002
Empirical Variance = 0.00196
Lambda=8, N=500
Empirical Mean = 7.9957
Theoretical Variance = 0.016
Empirical Variance = 0.01705
Density
Means of X

Example Two: "Real" Data

Back to the English Premier League!

```
> PL
```

	Rank	Team	GamesPlayed	Won	Drew	Lost	GoalsFor	GoalsAgainst	GoalDifference	Points
2	1	Liverpool	7	6	0	1	13	3	10	18
3	2	Manchester City	7	5	2	0	17	8	9	17
4	3	Arsenal	7	5	2	0	15	6	9	17
5	4	Chelsea	7	4	2	1	16	8	8	14
6	5	Aston Villa	7	4	2	1	12	9	3	14
7	6	Brighton & Hove Albion	7	3	3	1	13	10	3	12
8	7	Newcastle United	7	3	3	1	8	7	1	12
9	8	Fulham	7	3	2	2	10	8	2	11
10	9	Tottenham Hotspur	7	3	1	3	14	8	6	10
11	10	Nottingham Forest	7	2	4	1	7	6	1	10
12	11	Brentford	7	3	0	4	13	13	0	10
13	12	West Ham United	7	2	2	3	10	11	-3	8
14	13	Bournemouth	7	2	2	3	8	10	-2	8
15	14	Manchester United	7	2	2	3	5	8	-3	8
16	15	Leicester City	7	1	3	3	9	12	-3	6
17	16	Everton	7	1	2	4	7	15	-8	5
18	17	Ipswich Town	7	0	4	3	6	14	-8	4
19	18	Crystal Palace	7	0	3	4	5	10	-5	3
20	19	Southampton	7	0	1	6	4	15	-11	1
21	20	Wolverhampton Wanderers	7	0	1	6	9	21	-12	1

Premier League: Summary

```
> psych::describe(PL)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Rank*	1	20	10.50	5.92	10.5	10.50	7.41	1	20	19	0.00	-1.38	1.32
Team*	2	20	10.50	5.92	10.5	10.50	7.41	1	20	19	0.00	-1.38	1.32
GamesPlayed	3	20	7.00	0.00	7.0	7.00	0.00	7	7	0	NaN	NaN	0.00
Won	4	20	2.45	1.79	2.5	2.38	2.22	0	6	6	0.18	-1.02	0.40
Drew	5	20	2.05	1.10	2.0	2.06	1.48	0	4	4	-0.09	-0.56	0.25
Lost	6	20	2.50	1.76	3.0	2.38	2.22	0	6	6	0.41	-0.80	0.39
GoalsFor	7	20	10.05	3.89	9.5	9.94	5.19	4	17	13	0.16	-1.29	0.87
GoalsAgainst	8	20	10.10	4.05	9.5	9.81	2.97	3	21	18	0.79	0.51	0.91
GoalDifference	9	20	-0.15	6.63	0.5	0.06	6.67	-12	10	22	-0.12	-1.12	1.48
Points	10	20	9.45	5.11	10.0	9.50	5.93	1	18	17	-0.02	-1.09	1.14

Fitting Distributions To Data

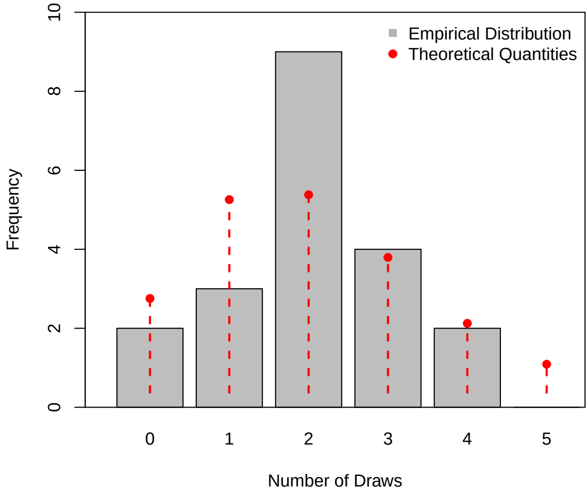
Useful commands / packages:

- `fitdistr` (in `MASS`; fits beta, Cauchy, chi-squared, exponential, gamma, geometric, log-normal, logistic, negative binomial, normal, Poisson, t, and weibull)
- `VGAM` has a lot of functions for fitting (many dofferent) distributions as well
- The `fitdistrplus` package extends `fitdistr` in some useful ways, as does `extraDistr`
- Visualization via `visualize`, `vistributions`
- See in general the insanely comprehensive CRAN Task View for *Distributions*

Fitting a Poisson Distribution

```
> library(MASS)
> PoisMean <- fitdistr(PL$Drew,"poisson")
> PoisMean
  lambda
  2.05
 (0.32)

> # Components:

> coef(PoisMean)
lambda
  2.05

> vcov(PoisMean)
       lambda
lambda  0.103

> # Note also:
>
> coef(PoisMean) / nrow(PL)
lambda
 0.102

> # and:
>
> (PoisMean$sd)^2
lambda
 0.103
```

Actual vs. Theoretical Draws with $\lambda = 2.05$

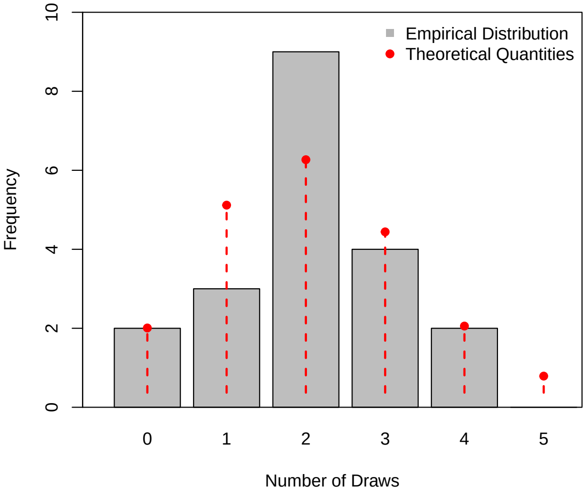


Fitting a Binomial Distribution

In this case, the better (read: more correct) distribution is the binomial with $N = 7$ (i.e., how many draws in seven matches):

```
> library(fitdistrplus)
> BinMean <- fitdist(PL$Drew,"binom",fix.arg=list(size=7))
> BinMean
Fitting of the distribution 'binom' by maximum likelihood
Parameters:
     estimate Std. Error
prob    0.293     0.0385
Fixed parameters:
     value
size     7
```

Actual vs. Theoretical Draws with $\hat{\pi} = 0.293$



Which One Fits "Better"?

For two distributions, how can we know which one is a better "fit" to the data?

This is tricky, because we often don't know the true distribution of the data...

Consider:

- Some data on X, which...
- ...is drawn from a distribution f, and...
- ...two "candidate" distributions, g_1 and g_2.

If we knew f, we could compare how different / wrong g_1 is in characterizing X, compared to g_2...

Akaike (1974): Compare how different / wrong g_1 is relative to g_2...

For a given model / distribution g, the *Akaike Information Criterion (AIC)* is:

$$AIC = 2k - 2\ln(\widehat{L_g})$$

where k is the number of parameters estimated in the model and $\widehat{L_g}$ is the estimated *likelihood* of the distribution / model g.

We'll talk more about likelihoods in the (near) future...

AIC Explained

Key points:

- AIC is a measure of the relative information of the model vis-a-vis the data...
- ...specifically, the information *loss* due to the lack of model "fit" to the data.
- Because log-likelihoods are negative, AIC will (almost) always be positive
- **Smaller values of AIC → better-fitting model / distribution**
- More specifically, for two models / distributions g_1 and g_2, the *relative likelihood*:

$$RL = \exp\left(\frac{AIC_{g_1} - AIC_{g_2}}{2}\right)$$

...is proportional to the probability that g_2 is more likely to minimize the information loss than g_1

A Variant: BIC

An alternative to the AIC is the "Bayesian Information Criterion ("BIC")[2]

$$BIC = k\ln(N) - 2\ln(\widehat{L_g})$$

- BIC is also a measure of the relative information of the model vis-a-vis the data
- Once again, smaller values of BIC $\rightarrow$ better-fitting model / distribution
- Both AIC and BIC are useful; a comparison + discussion is here

[2] Also sometimes called the "Schwarz Information Criterion" ("SIC"), after Schwarz (1978).

Simulation Example

To illustrate, we'll:

1. Draw a sample of size N from a standard Normal $[N(0,1)]$ distribution
2. Fit three distributions to the resulting data:
   - A normal distribution (estimating $\hat{\mu}$ and $\hat{\sigma}^2$)
   - A t distribution (estimating $\hat{\nu}$, the degrees of freedom parameter)
   - A *LaPlace distribution*, which has density:

   $$f(X) \equiv \Pr(X = x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

   (estimating $\hat{\mu}$ and $\hat{b}$)
3. Compare their AIC and BIC statistics...

Simulation: AIC and BIC

```
> set.seed(7222009)
> X<-rnorm(200,0,1)                     # 200 observations from N(0,1)
>
> # Define the LaPlace density:
>
> dlaplace <- function(x, mu=0, b=1){
>   if(b<=0) return(NA)
>   exp(-abs(x-mu)/b) / (2*b)
> }
>
> # Fit using -fitdist-:
>
> NormFit<-fitdist(X,"norm")             # Normal
> TFit<-fitdist(X,"t",start=list(df=3))  # t
> LaPlaceFit<-fitdist(X,dlaplace,        # LaPlace
>             start=list(b=1,mu=0))
>
> # Display differences:
>
> Dists<-list("Normal"=NormFit,"t"=TFit,"LaPlace"=LaPlaceFit)
> aictab(Dists)

Model selection based on AICc:

        K AICc Delta_AICc AICcWt Cum.Wt   LL
t       1  573       0.00   0.64   0.64 -286
Normal  2  574       1.18   0.36   1.00 -285
LaPlace 2  605      31.46   0.00   1.00 -300
```

AIC Comparisons

Comparing Normal to LaPlace:

$$\begin{aligned} RL &= \exp\left(\frac{574 - 604}{2}\right) \\ &= \exp(-15) \\ &= 0.0000003 \end{aligned}$$

Implication: The LaPlace distribution is 0.0000003 times as probable as the Normal distribution to minimize the information loss.

Comparing t to Normal:

$$\begin{aligned} RL &= \exp\left(\frac{573 - 574}{2}\right) \\ &= \exp(-0.50) \\ &= 0.607 \end{aligned}$$

Implication: The Normal distribution is 0.607 times as probable as the t distribution to minimize the information loss.

Premier League Draws: AIC Comparison

```
> BinMean<-fitdist(PL$Drew,"binom",fix.arg=list(size=7))
> PoisMean<-fitdist(PL$Drew,"pois")

> print(c(round(BinMean$bic,3),round(PoisMean$bic,3)))    # BICs
[1] 63.2 65.7

> # Compare AIC + BIC:
>
> Draws<-list("Poisson"=PoisMean,"Binomial"=BinMean)
> aictab(Draws)

Model selection based on AICc:

         K AICc Delta_AICc AICcWt Cum.Wt    LL
Binomial 1 62.4       0.00   0.78   0.78 -30.1
Poisson  1 64.9       2.49   0.22   1.00 -31.3
```

Interpretation:

$$
\begin{aligned}
RL &= \exp\left(\frac{62.4 - 64.9}{2}\right) \\
&= \exp(-1.25) \\
&= 0.287
\end{aligned}
$$

Implication: The Poisson distribution is 0.287 times as probable as the binomial distribution to minimize the information loss.