The Pythagorean Won-Loss Formula in Baseball

An Introduction to Statistics and Modeling

Steven J. Miller

Brown University

Providence, December 7^{th}, 2005

http://www.math.brown.edu/∼sjmiller

Probability Review

Probability density:

- $p(x) \geq 0$;
- $\int_{-\infty}^{\infty} p(x) \mathrm{d}x = 1$;
- X random variable with density $p(x)$: $\mathrm{Prob}\left(X \in [a, b]\right) = \int_a^b p(x) \mathrm{d}x$.

Mean (average value) $\mu = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x$.

Variance (how spread out) $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) \mathrm{d}x$.

Independence: two random variables are independent if knowledge of one does not give knowledge of the other.

Numerical Observation: Pythagorean Won-Loss Formula

Parameters:

- $\mathrm{RS}_{\mathrm{obs}}$: average number of runs scored per game;

- $\mathrm{RA}_{\mathrm{obs}}$: average number of runs allowed per game;

- γ: some parameter, constant for a sport.

Bill James' Won-Loss Formula (NUMERICAL Observation):

$$\mathrm{Won-Loss\ Percentage} = \frac{{\mathrm{RS}_{\mathrm{obs}}}^{\gamma}}{{\mathrm{RS}_{\mathrm{obs}}}^{\gamma} + {\mathrm{RA}_{\mathrm{obs}}}^{\gamma}}$$

For baseball: γ originally taken as 2.
Numerical studies show best γ is about 1.82.

Modeling the Real World

Guidelines for Modeling:

- Model should capture key features of the system;
- Model should be mathematically tractable (solvable).

In general these are conflicting goals. How should we try and model baseball games?

Possible Model:

- Runs Scored and Runs Allowed independent random variables;
- $f_{\rm RS}(x)$, $g_{\rm RA}(y)$: probability density functions for runs scored (allowed).
- Reduced to calculating

$$\int_x \left[\int_{y \le x} f_{\rm RS}(x) g_{\rm RA}(y) \mathrm{d}y\right] \mathrm{d}x \quad \text{or} \quad \sum_i \left[\sum_{j<i} f_{\rm RS}(i) g_{\rm RA}(j)\right].$$

Problems with the Model

Reduced to calculating

$$\int_x \left[\int_{y \le x} f_{\mathrm{RS}}(x) g_{\mathrm{RA}}(y) \mathrm{d}y \right] \mathrm{d}x \quad \text{or} \quad \sum_i \left[\sum_{j<i} f_{\mathrm{RS}}(i) g_{\mathrm{RA}}(j) \right].$$

Problems with the model:

- Can the integral (or sum) be completed in closed form?
- Are the runs scored and allowed independent random variables?
- What are f_{RS} and g_{RA}?

Three Parameter Weibull

Weibull distribution:

$$f(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$

- α: scale (meters versus centimeters);
- β: origin (translation, zero point);
- γ: shape (behavior near β and at infinity).

Various values give different shapes, but can we find α, β, γ such that it fits observed data? Is the Weibull theoretically tractable?

Weibull Integrations

Let $f(x;\alpha,\beta,\gamma)$ be the probability density of a Weibull(α,β,γ):

$$f(x;\alpha,\beta,\gamma) \;=\; \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$

For $s \in \mathbb{C}$ with the real part of s greater than 0, recall the Γ-function:

$$\Gamma(s) \;=\; \int_0^\infty e^{-u} u^{s-1} \mathrm{d}u \;=\; \int_0^\infty e^{-u} u^{s} \frac{\mathrm{d}u}{u}.$$

Let $\mu_{\alpha,\beta,\gamma}$ denote the mean of $f(x;\alpha,\beta,\gamma)$.

Weibull Integrations (Continued)

$$
\begin{aligned}
\mu_{\alpha,\beta,\gamma} &= \int_\beta^\infty x \cdot \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma}\mathrm{d}x \\
&= \int_\beta^\infty \alpha\frac{x-\beta}{\alpha}\cdot\frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma}\mathrm{d}x \;+\; \beta.
\end{aligned}
$$

Change variables: $u = \left(\frac{x-\beta}{\alpha}\right)^\gamma$. Then $\mathrm{d}u = \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1}\mathrm{d}x$ and

$$
\begin{aligned}
\mu_{\alpha,\beta,\gamma} &= \int_0^\infty \alpha u^{\gamma^{-1}}\cdot e^{-u}\mathrm{d}u \;+\; \beta \\
&= \alpha\int_0^\infty e^{-u}u^{1+\gamma^{-1}}\frac{\mathrm{d}u}{u} \;+\; \beta \\
&= \alpha\Gamma(1+\gamma^{-1}) \;+\; \beta.
\end{aligned}
$$

A similar calculation determines the variance.

Derivation of the Pythagorean Won-Loss Formula

Theorem: Pythagorean Won-Loss Formula: *Let the runs scored and allowed per game be two independent random variables drawn from Weibull distributions* $(\alpha_{\rm RS}, \beta, \gamma)$ *and* $(\alpha_{\rm RA}, \beta, \gamma)$*;* $\alpha_{\rm RS}$ *and* $\alpha_{\rm RA}$ *are chosen so that the means are* RS *and* RA*. If* $\gamma > 0$ *then*

$$\text{Won-Loss Percentage}({\rm RS}, {\rm RA}, \beta, \gamma) \ = \ \frac{({\rm RS}-\beta)^\gamma}{({\rm RS}-\beta)^\gamma + ({\rm RA}-\beta)^\gamma}.$$

Proof. Let X and Y be independent random variables with Weibull distributions $(\alpha_{\rm RS}, \beta, \gamma)$ and $(\alpha_{\rm RA}, \beta, \gamma)$ respectively.

$$\alpha_{\rm RS} \ = \ \frac{{\rm RS}-\beta}{\Gamma(1+\gamma^{-1})}, \qquad \alpha_{\rm RA} \ = \ \frac{{\rm RA}-\beta}{\Gamma(1+\gamma^{-1})}.$$

We need only calculate the probability that X exceeds Y. We use the integral of a probability density is 1.

$$
\begin{aligned}
\text{Prob}(X > Y) &= \int_{x=\beta}^{\infty} \int_{y=\beta}^{x} f(x; \alpha_{\mathrm{RS}}, \beta, \gamma) f(y; \alpha_{\mathrm{RA}}, \beta, \gamma) \mathrm{d}y \, \mathrm{d}x \\
&= \int_{x=\beta}^{\infty} \int_{y=\beta}^{x} \frac{\gamma}{\alpha_{\mathrm{RS}}} \left(\frac{x-\beta}{\alpha_{RS}} \right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha_{\mathrm{RS}}}\right)^{\gamma}} \frac{\gamma}{\alpha_{\mathrm{RA}}} \left(\frac{y-\beta}{\alpha_{\mathrm{RA}}} \right)^{\gamma-1} e^{-\left(\frac{y-\beta}{\alpha_{\mathrm{RA}}}\right)^{\gamma}} \mathrm{d}y \\
&= \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\mathrm{RS}}} \left(\frac{x}{\alpha_{RS}} \right)^{\gamma-1} e^{-\left(\frac{x}{\alpha_{\mathrm{RS}}}\right)^{\gamma}} \left[\int_{y=0}^{x} \frac{\gamma}{\alpha_{\mathrm{RA}}} \left(\frac{y}{\alpha_{\mathrm{RA}}} \right)^{\gamma-1} e^{-\left(\frac{y}{\alpha_{\mathrm{RA}}}\right)^{\gamma}} \mathrm{d}y \right] \\
&= \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\mathrm{RS}}} \left(\frac{x}{\alpha_{RS}} \right)^{\gamma-1} e^{-(x/\alpha_{\mathrm{RS}})^{\gamma}} \left[1 - e^{-(x/\alpha_{\mathrm{RA}})^{\gamma}} \right] \mathrm{d}x \\
&= 1 - \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\mathrm{RS}}} \left(\frac{x}{\alpha_{RS}} \right)^{\gamma-1} e^{-(x/\alpha)^{\gamma}} \mathrm{d}x,
\end{aligned}
$$

where we have set

$$
\frac{1}{\alpha^{\gamma}} = \frac{1}{\alpha_{\mathrm{RS}}^{\gamma}} + \frac{1}{\alpha_{\mathrm{RA}}^{\gamma}} = \frac{\alpha_{\mathrm{RS}}^{\gamma} + \alpha_{\mathrm{RA}}^{\gamma}}{\alpha_{\mathrm{RS}}^{\gamma} \alpha_{\mathrm{RA}}^{\gamma}}.
$$

$$\begin{aligned}
\text{Prob}(X > Y) &= 1 - \frac{\alpha^\gamma}{\alpha_{\text{RS}}^\gamma} \int_0^\infty \frac{\gamma}{\alpha} \left(\frac{x}{\alpha}\right)^{\gamma-1} e^{(x/\alpha)^\gamma} \mathrm{d}x \\
&= 1 - \frac{\alpha^\gamma}{\alpha_{\text{RS}}^\gamma} \\
&= 1 - \frac{1}{\alpha_{\text{RS}}^\gamma} \frac{\alpha_{\text{RS}}^\gamma \alpha_{\text{RA}}^\gamma}{\alpha_{\text{RS}}^\gamma + \alpha_{\text{RA}}^\gamma} \\
&= \frac{\alpha_{\text{RS}}^\gamma}{\alpha_{\text{RS}}^\gamma + \alpha_{\text{RA}}^\gamma}.
\end{aligned}$$

We substitute the relations for α_{RS} and α_{RA} and find that

$$\text{Prob}(X > Y) = \frac{(\text{RS} - \beta)^\gamma}{(\text{RS} - \beta)^\gamma + (\text{RA} - \beta)^\gamma}.$$

Note $\text{RS} - \beta$ estimates RS_{obs}, $\text{RA} - \beta$ estimates RA_{obs}. $\square$

Best Fit Weibulls to Data: Method of Least Squares

Minimized the sum of squares of the error from the runs scored data plus the sum of squares of the error from the runs allowed data.

- $\mathrm{Bin}(k)$ is the k^{th} bin;
- $\mathrm{RS}_{\mathrm{obs}}(k)$ (resp. $\mathrm{RA}_{\mathrm{obs}}(k)$) the observed number of games with the number of runs scored (allowed) in $\mathrm{Bin}(k)$;
- $A(\alpha, \beta, \gamma, k)$ the area under the Weibull with parameters (α, β, γ) in $\mathrm{Bin}(k)$.

Find the values of $(\alpha_{\mathrm{RS}}, \alpha_{\mathrm{RA}}, \gamma)$ that minimize

$$\sum_{k=1}^{\#\mathrm{Bins}} \left(\mathrm{RS}_{\mathrm{obs}}(k) - \#\mathrm{Games} \cdot A(\alpha_{\mathrm{RS}}, -.5, \gamma, k)\right)^2 + \sum_{k=1}^{\#\mathrm{Bins}} \left(\mathrm{RA}_{\mathrm{obs}}(k) - \#\mathrm{Games} \cdot A(\alpha_{\mathrm{RA}}, -.5, \gamma, k)\right)^2 .$$

Best Fit Weibulls to Data: Method of Maximum Likelihood

The likelihood function depends on: $\alpha_{\rm RS}, \alpha_{\rm RA}, \beta = -.5, \gamma$.

Let $A(\alpha, -.5, \gamma, k)$ denote the area in $\mathrm{Bin}(k)$ of the Weibull with parameters $\alpha, -.5, \gamma$. The sample likelihood function $L(\alpha_{\rm RS}, \alpha_{\rm RA}, -.5, \gamma)$ is

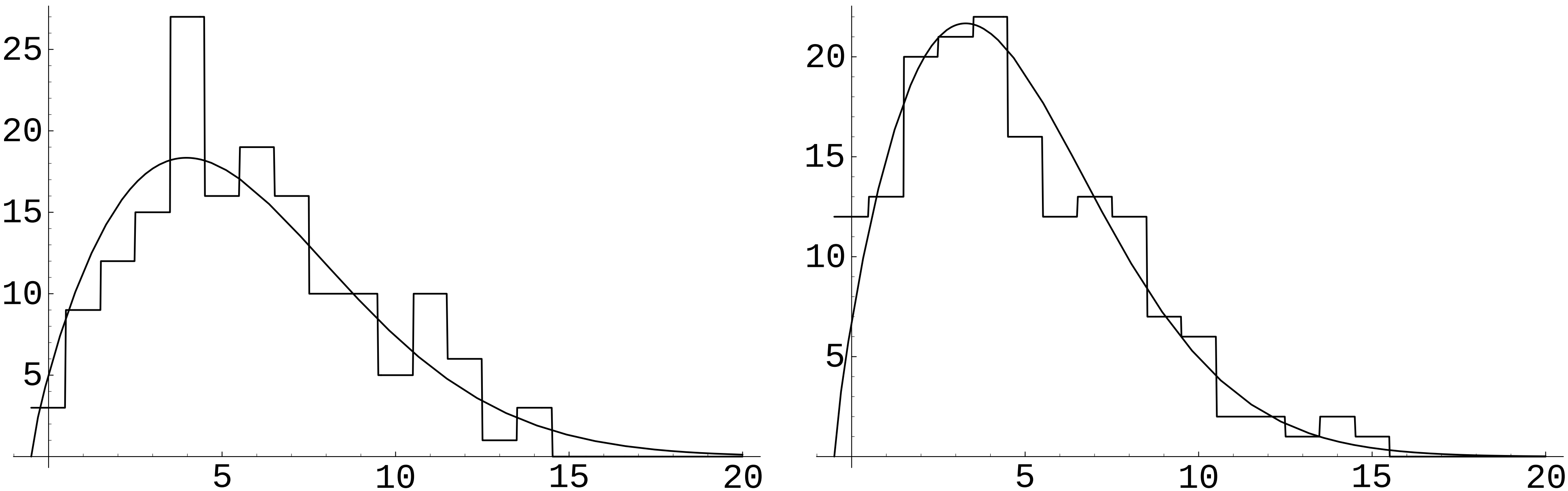
$$\binom{\#\text{Games}}{\text{RS}_{\rm obs}(1), \dots, \text{RS}_{\rm obs}(\#\text{Bins})} \prod_{k=1}^{\#\text{Bins}} A(\alpha_{\rm RS}, -.5, \gamma, k)^{\text{RS}_{\rm obs}(k)}$$

$$\cdot \binom{\#\text{Games}}{\text{RA}_{\rm obs}(1), \dots, \text{RA}_{\rm obs}(\#\text{Bins})} \prod_{k=1}^{\#\text{Bins}} A(\alpha_{\rm RA}, -.5, \gamma, k)^{\text{RA}_{\rm obs}(k)}.$$

For each team we find the values of the parameters $\alpha_{\rm RS}$, $\alpha_{\rm RA}$ and γ that maximize the likelihood. Computationally, it is equivalent to maximize the logarithm of the likelihood, and we may ignore the multinomial coefficients are they are independent of the parameters.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

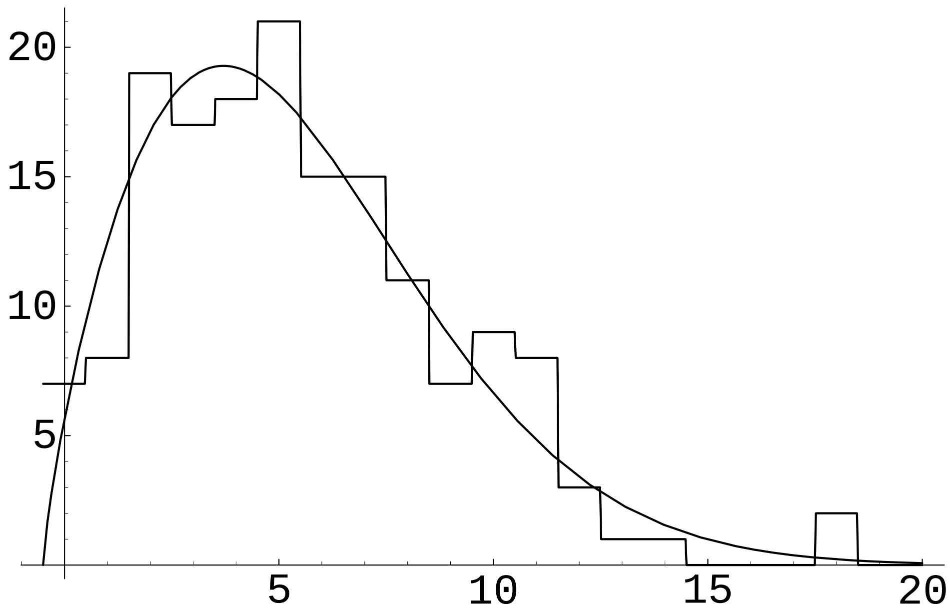
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Boston Red Sox

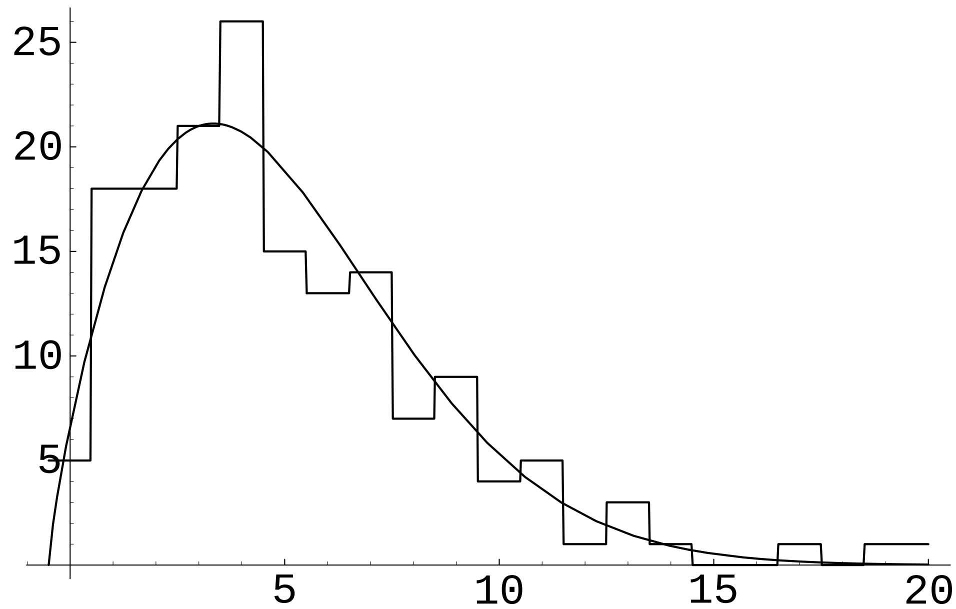


Using as bins

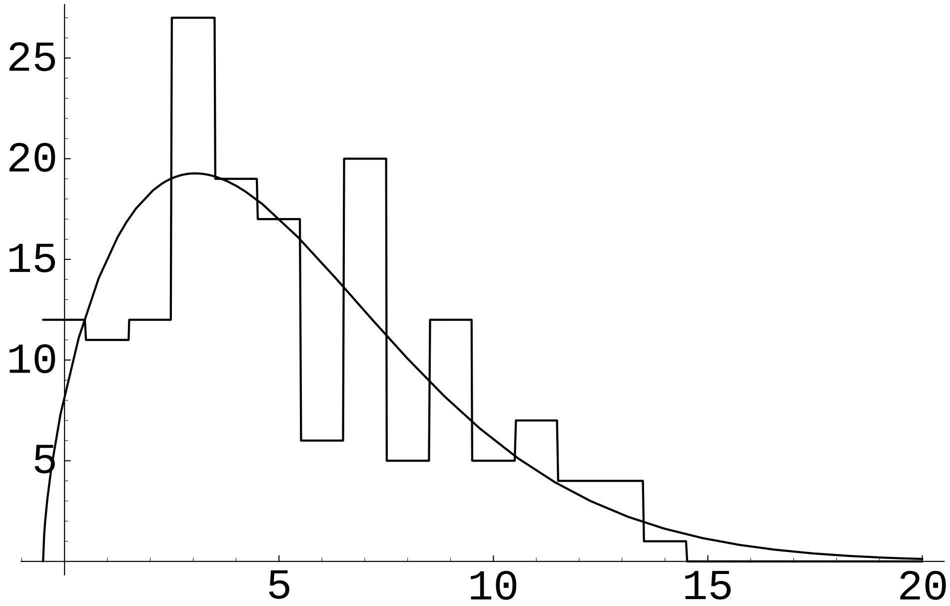
$$[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty).$$

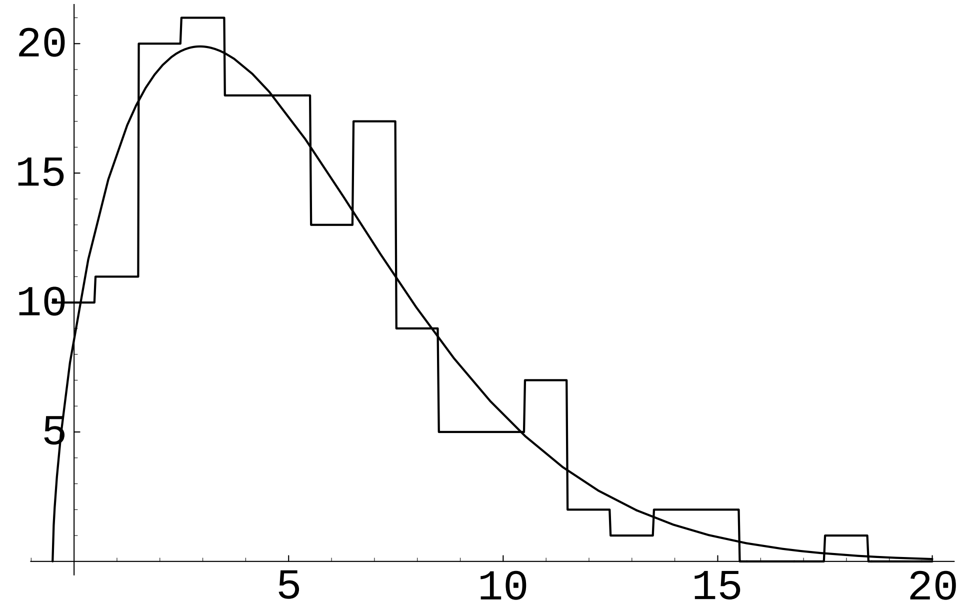
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the New York Yankees



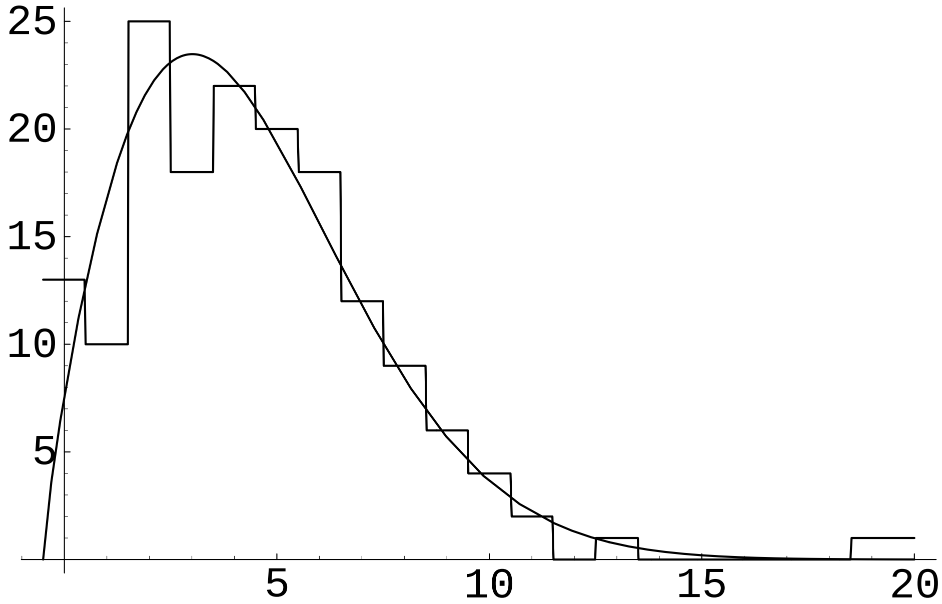


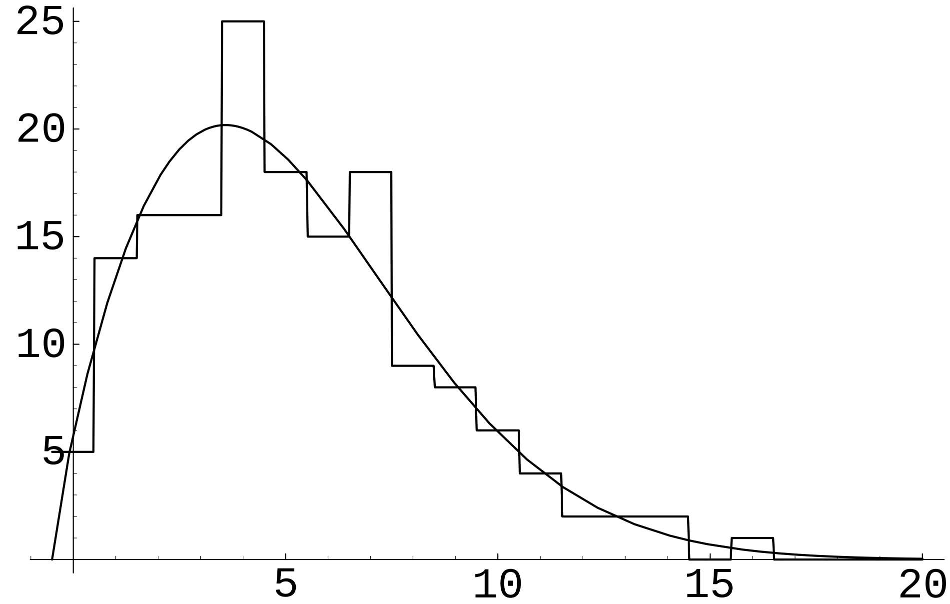
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Baltimore Orioles



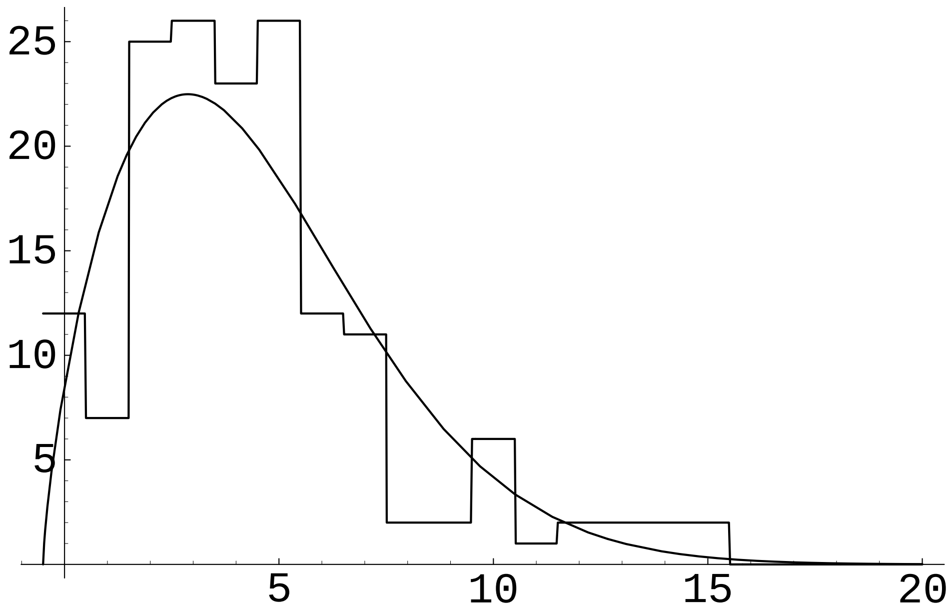


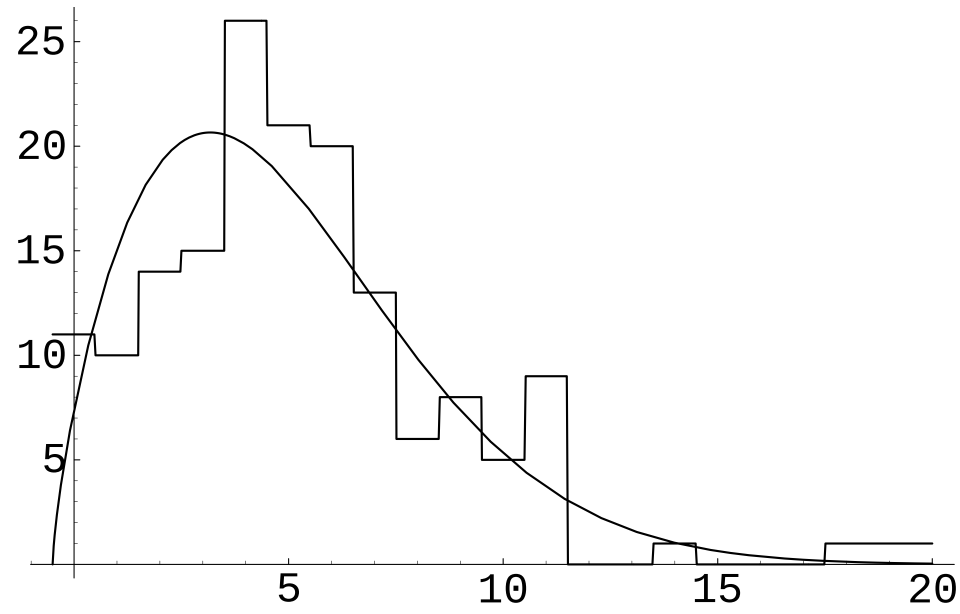
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Tampa Bay Devil Rays



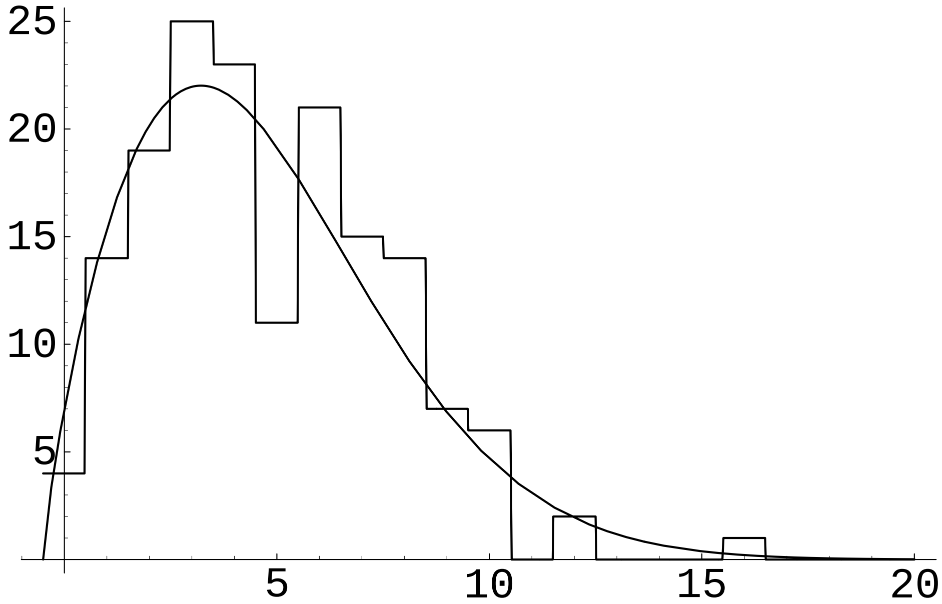


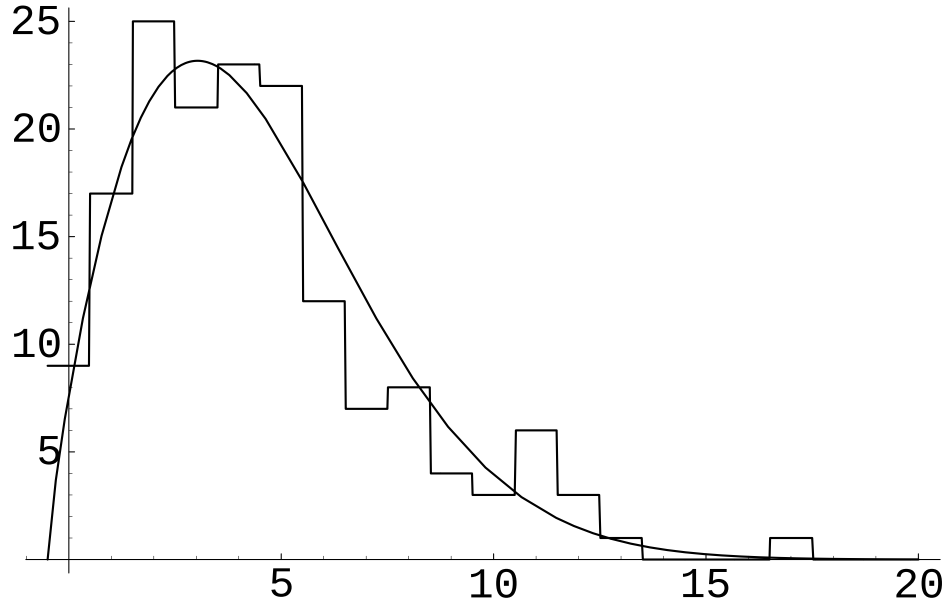
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Toronto Blue Jays



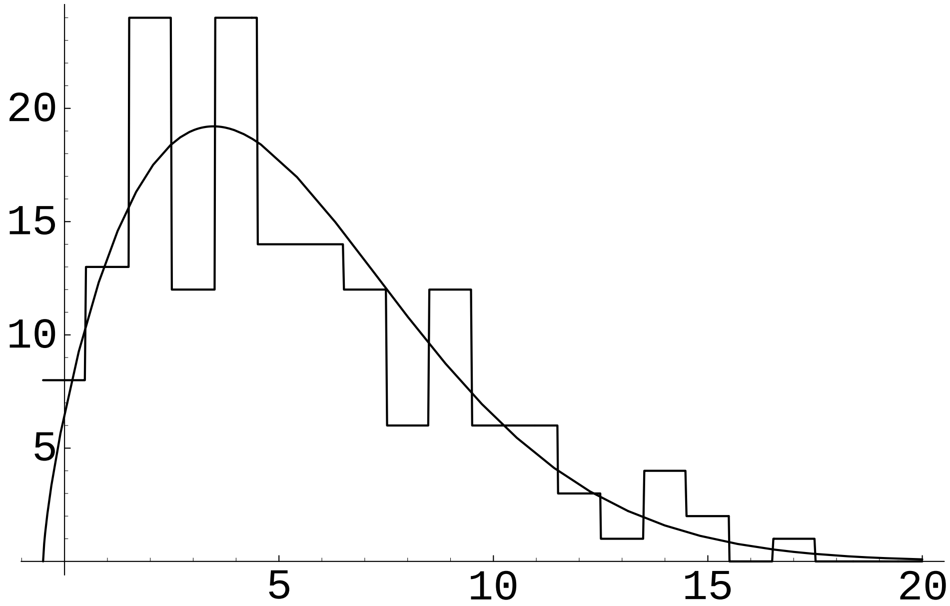


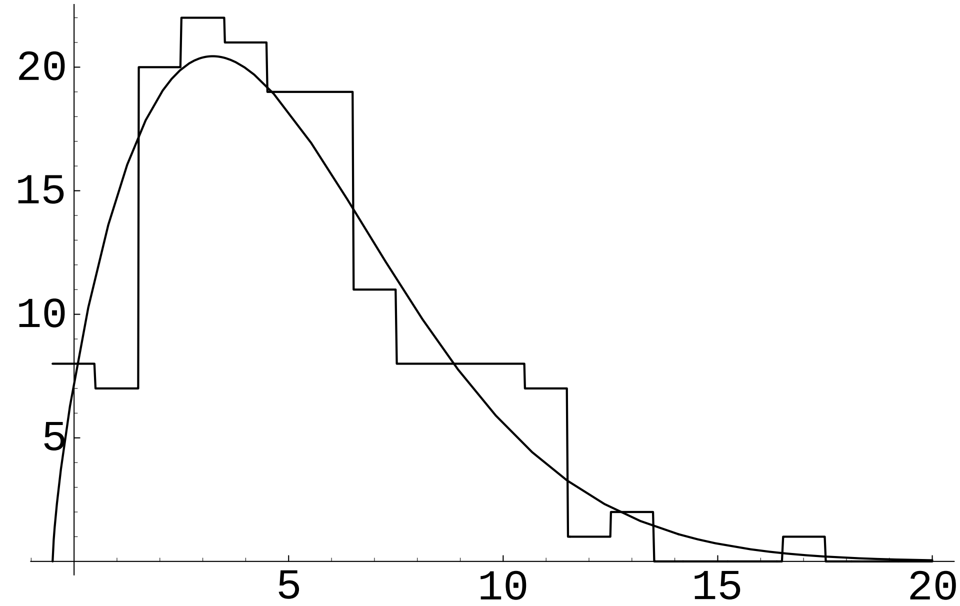
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Minnesota Twins



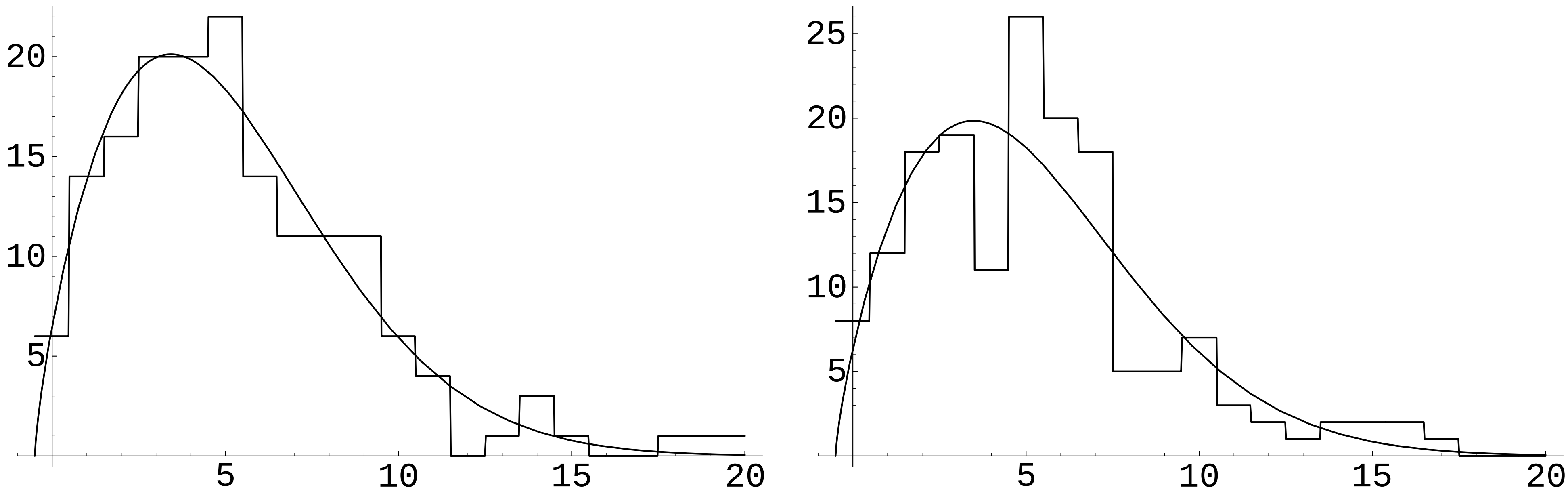


Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Chicago White Sox

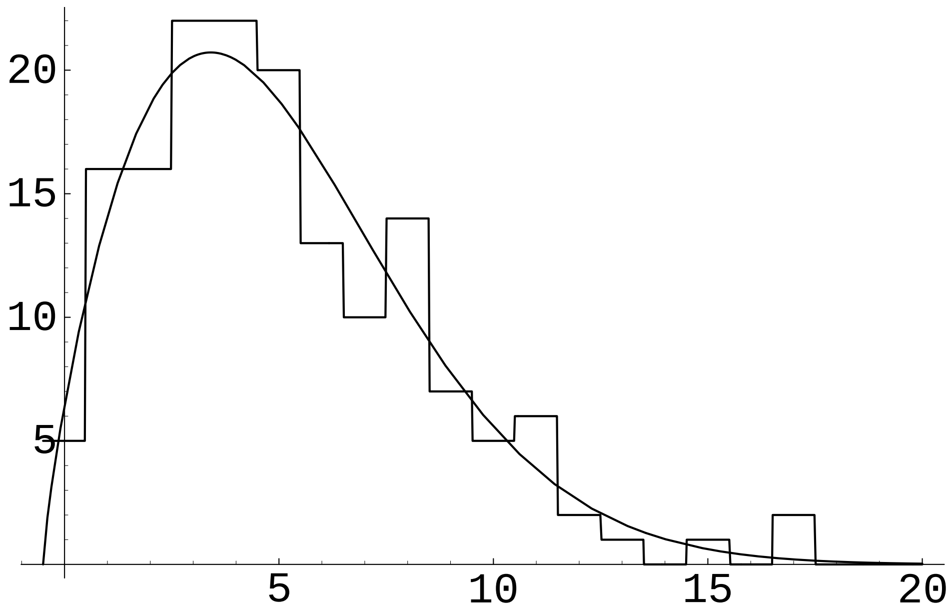


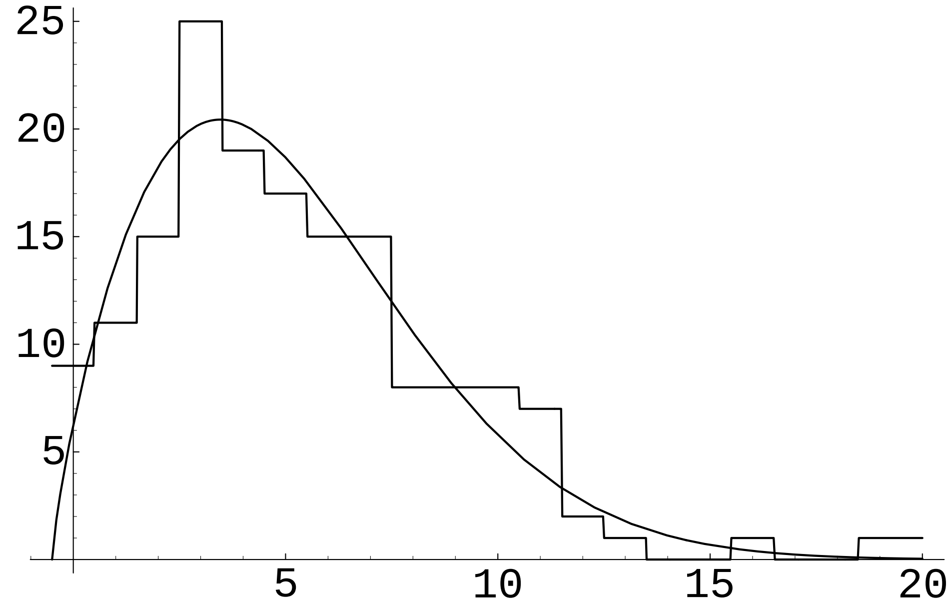


Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Cleveland Indians

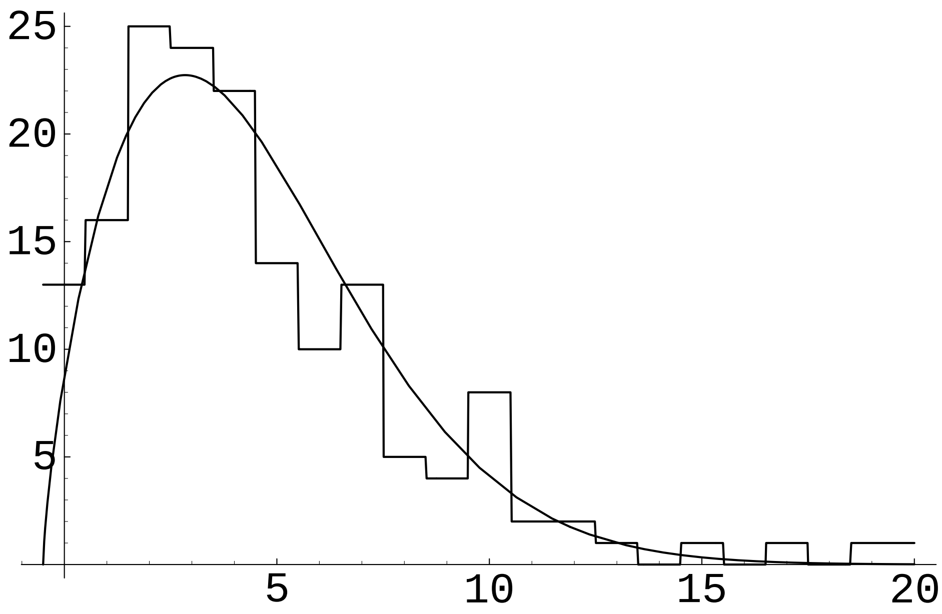


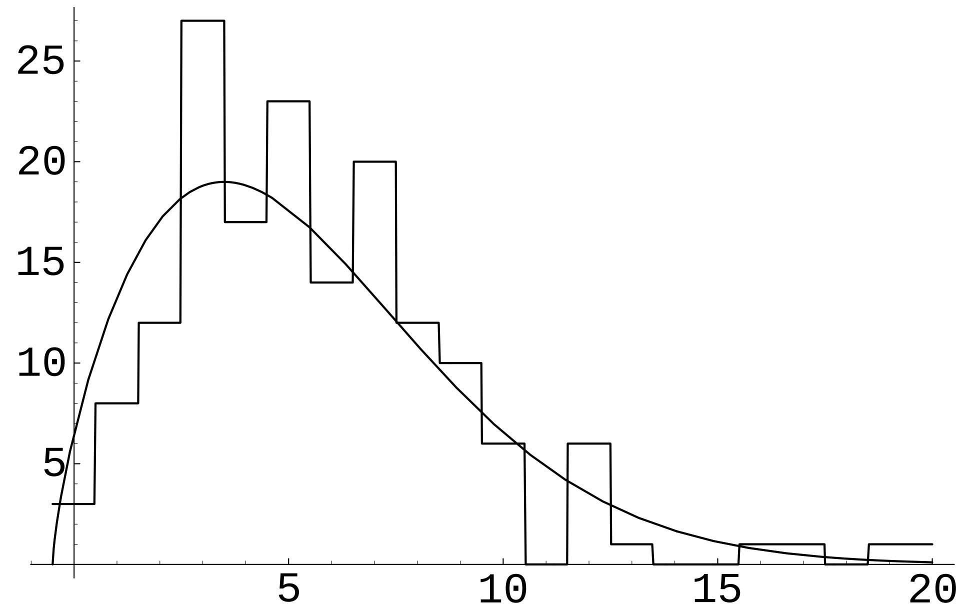
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Detroit Tigers



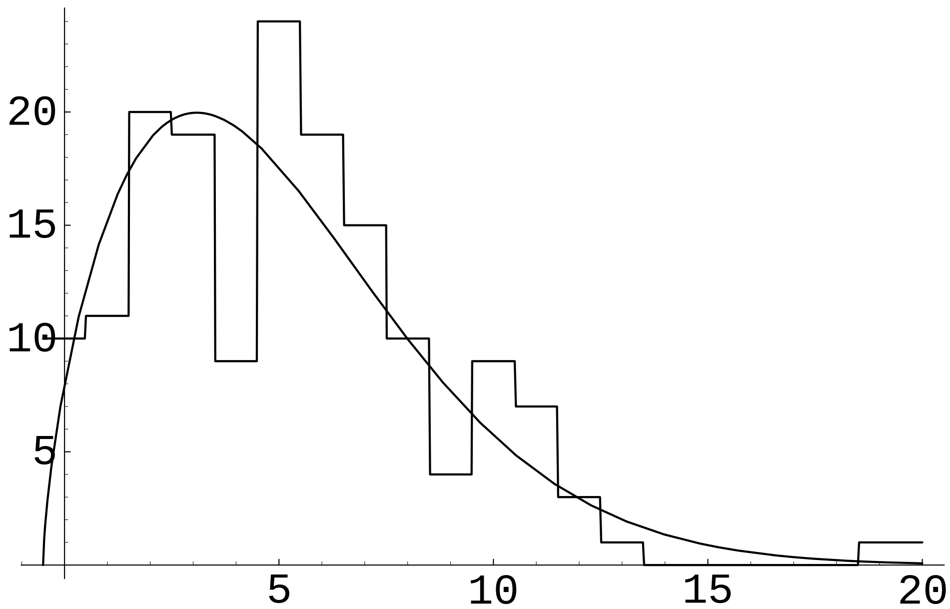


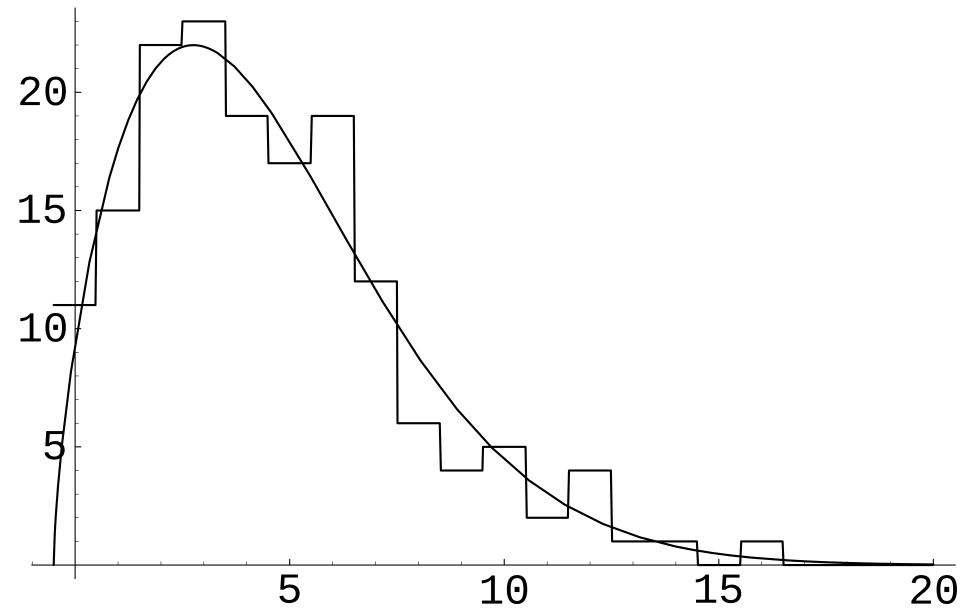
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Kansas City Royals



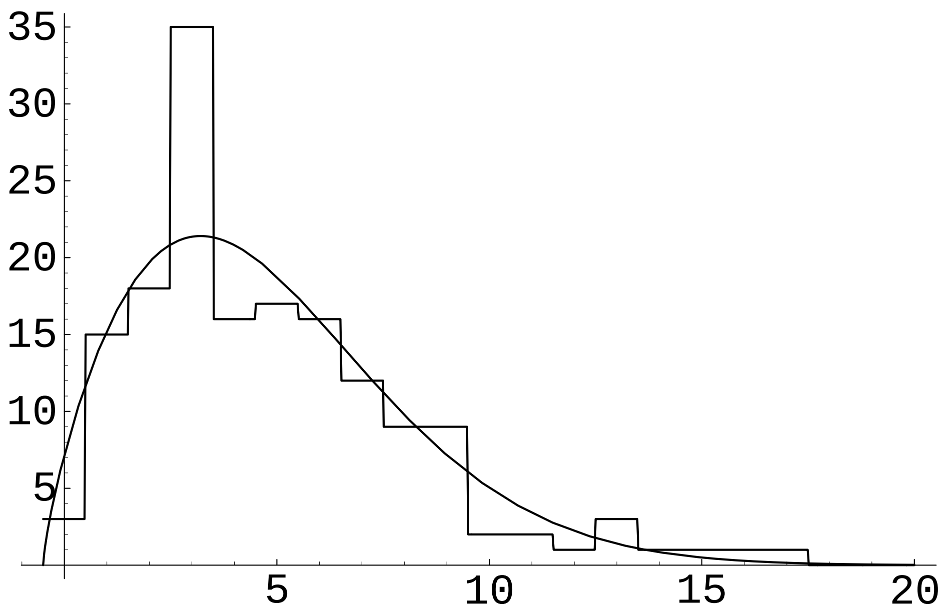


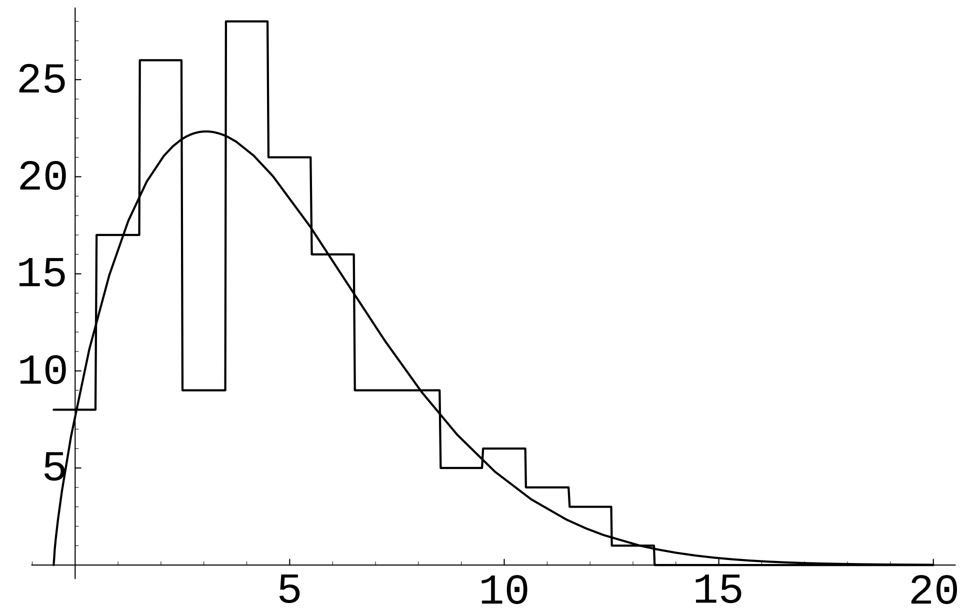
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Los Angeles Angels



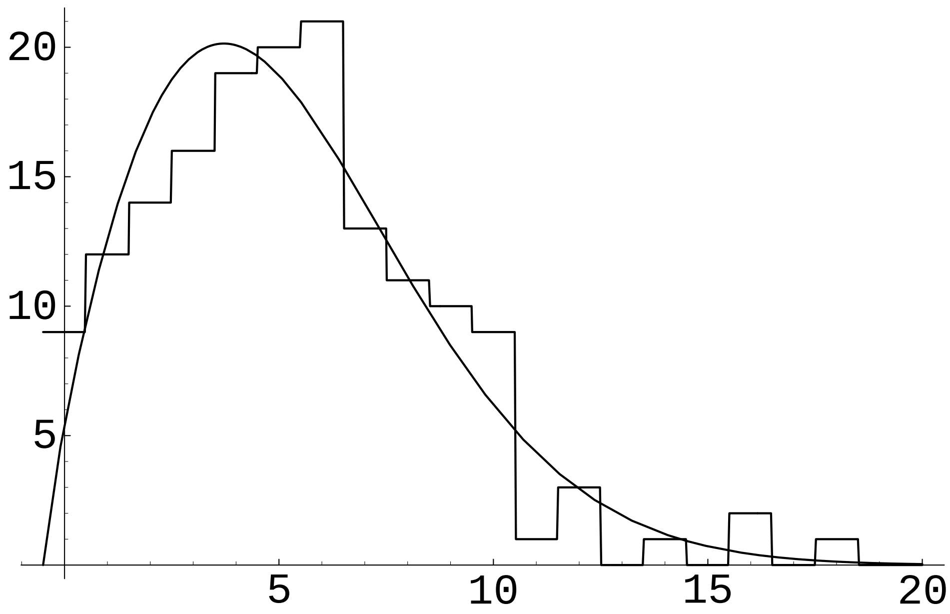


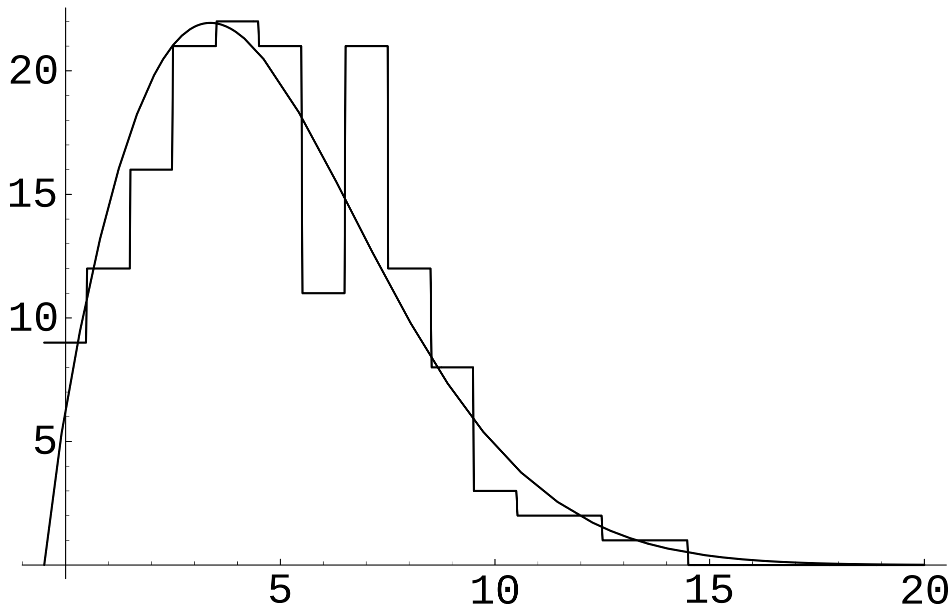
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Oakland Athletics



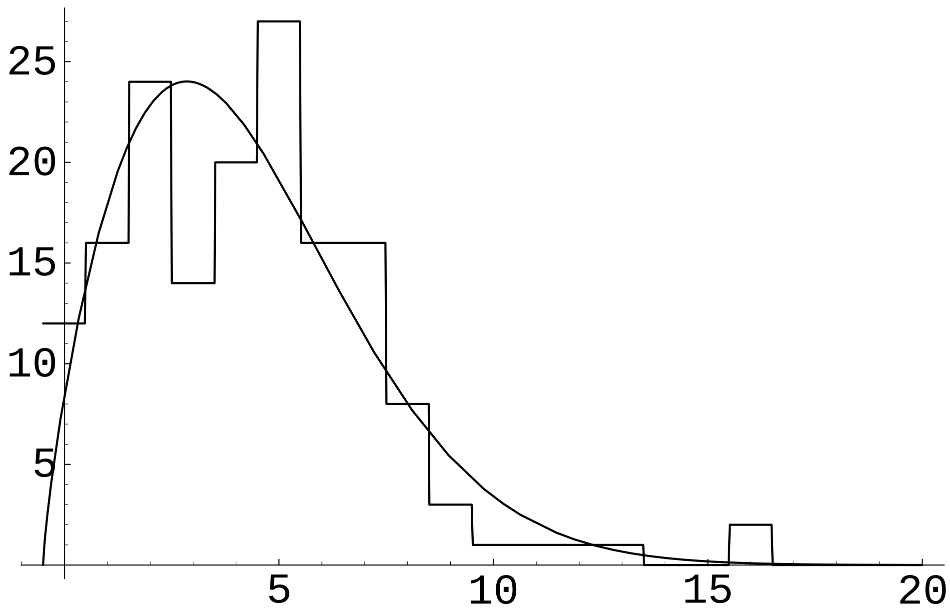


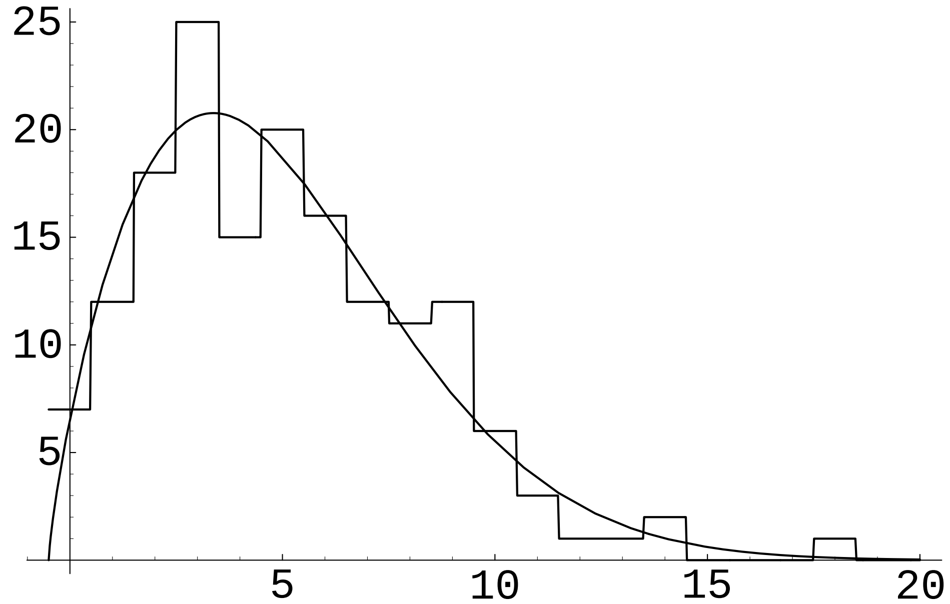
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Texas Rangers





Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Seattle Mariners





Best Fit Weibulls to Data (Continued)

The fits *look* good, but are they? Do χ^2-tests:

- Let $\mathrm{Bin}(k)$ denote the k^{th} bin.
- $O_{r,c}$: the observed number of games where the team's runs scored is in $\mathrm{Bin}(r)$ and the runs allowed are in $\mathrm{Bin}(c)$.
- $E_{r,c} = \frac{\sum_{c'} O_{r,c'} \cdot \sum_{r'} O_{r',c}}{\#\mathrm{Games}}$ is the expected frequency of cell (r, c).
- Then
$$\sum_{r=1}^{\#\mathrm{Rows}} \sum_{c=1}^{\#\mathrm{Columns}} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$
is a χ^2 distribution with $(\#\mathrm{Rows} - 1)(\#\mathrm{Columns} - 1)$ degrees of freedom.

Best Fit Weibulls to Data (Continued)

For independence of runs scored and allowed, use bins

$$[0,1) \cup [1,2) \cup [2,3) \cup \cdots \cup [8,9) \cup [9,10) \cup [10,11) \cup [11,\infty).$$

Have an $r \times c$ contingency table (with $r = c = 12$); however, there are *structural zeros* (runs scored and allowed per game can never be equal). (Essentially) $O_{r,r} = 0$ for all r. We use the iterative fitting procedure to obtain maximum likelihood estimators for the $E_{r,c}$, the expected frequency of cell (r,c) under the assumption that, given that the runs scored and allowed are distinct, the runs scored and allowed are independent.

For $1 \le r, c \le 12$, let $E_{r,c}^{(0)} = 1$ if $r \neq c$ and 0 if $r = c$. Set

$$X_{r,+} = \sum_{c=1}^{12} O_{r,c}, \quad X_{+,c} = \sum_{r=1}^{12} O_{r,c}.$$

Then

$$E_{r,c}^{(\ell)} = \begin{cases} E_{r,c}^{(\ell-1)} X_{r,+} \;/\; \sum_{c=1}^{12} E_{r,c}^{(\ell-1)} & \text{if } \ell \text{ is odd} \\ \\ E_{r,c}^{(\ell-1)} X_{+,c} \;/\; \sum_{r=1}^{12} E_{r,c}^{(\ell-1)} & \text{if } \ell \text{ is even,} \end{cases}$$

and

$$E_{r,c} = \lim_{\ell \to \infty} E_{r,c}^{(\ell)};$$

the iterations converge very quickly. (If we had a complete two-dimensional contingency table, then the iteration reduces to the standard values, namely $E_{r,c} = \sum_{c'} O_{r,c'} \cdot \sum_{r'} O_{r',c} \;/\; \#\mathrm{Games}$.). Note

$$\sum_{r=1}^{12} \sum_{\substack{c=1 \\ c \neq r}}^{12} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

is approximately a χ^2 distribution with $(12-1)^2 - 12 = 109$ degrees of freedom. The corresponding critical thresholds are 134.4 (at the 95% level) and 146.3 (at the 99% level).

Team	RS+RA $\chi 2$: 20 d.f.	Indep $\chi 2$: 109 d.f
Boston Red Sox	15.63	83.19
New York Yankees	12.60	129.13
Baltimore Orioles	29.11	116.88
Tampa Bay Devil Rays	13.67	111.08
Toronto Blue Jays	41.18	100.11
Minnesota Twins	17.46	97.93
Chicago White Sox	22.51	153.07
Cleveland Indians	17.88	107.14
Detroit Tigers	12.50	131.27
Kansas City Royals	28.18	111.45
Los Angeles Angels	23.19	125.13
Oakland Athletics	30.22	133.72
Texas Rangers	16.57	111.96
Seattle Mariners	21.57	141.00

20 d.f.: 31.41 (at the 95% level) and 37.57 (at the 99% level).
109 d.f.: 134.4 (at the 95% level) and 146.3 (at the 99% level).

Bonferroni Adjustment:
20 d.f.: 41.14 (at the 95% level) and 46.38 (at the 99% level).
109 d.f.: 152.9 (at the 95% level) and 162.2 (at the 99% level).

Testing the Model: Data from Method of Maximum Likelihood

Team	Obs Wins	Pred Wins	ObsPerc	PredPerc	GamesDiff	
Boston Red Sox	98	93.0	0.605	0.574	5.03	1
New York Yankees	101	87.5	0.623	0.540	13.49	1
Baltimore Orioles	78	83.1	0.481	0.513	−5.08	1
Tampa Bay Devil Rays	70	69.6	0.435	0.432	0.38	1
Toronto Blue Jays	67	74.6	0.416	0.464	−7.65	1
Minnesota Twins	92	84.7	0.568	0.523	7.31	1
Chicago White Sox	83	85.3	0.512	0.527	−2.33	1
Cleveland Indians	80	80.0	0.494	0.494	0.	1
Detroit Tigers	72	80.0	0.444	0.494	−8.02	1
Kansas City Royals	58	68.7	0.358	0.424	−10.65	1
Los Angeles Angels	92	87.5	0.568	0.540	4.53	1
Oakland Athletics	91	84.0	0.562	0.519	6.99	1
Texas Rangers	89	87.3	0.549	0.539	1.71	1
Seattle Mariners	63	70.7	0.389	0.436	−7.66	1

γ: mean $= 1.74$, standard deviation $= .06$, median $= 1.76$;
close to numerically observed value of 1.82.

The mean number of the difference between observed and predicted wins was $-.13$ with a standard deviation of 7.11 (and a median of 0.19).
If we consider just the absolute value of the difference then we have a mean of 5.77 with a standard deviation of 3.85 (and a median of 6.04).

Conclusions

- Can find parameters such that the Weibulls are good fits to the data;
- The runs scored and allowed per game are statistically independent;
- The Pythagorean Won-Loss Formula is a consequence of our model;
- Our best value of γ of about 1.74 is close to the observed best 1.82.

Future Work

- Micro-analysis: runs scored and allowed are not entirely independent (big lead, close game), run production smaller for inter-league games in NL parks, et cetera.

- What about other sports? Does the same model work? How does γ depend on the sport?

- Are there other probability distributions that give integrals which can be determined in closed form?

Appendix: Proof of Central Limit Theorem: Notation

Convolution of f and g:

$$h(y) = \int_{\mathbb{R}} f(x)g(y-x)\mathrm{d}x = \int_{\mathbb{R}} f(x-y)g(x)\mathrm{d}x.$$

X_1 and X_2 independent random variables with probability density p.

$$\mathrm{Prob}(X_i \in [x, x+\Delta x]) = \int_x^{x+\Delta x} p(t)\mathrm{d}t \approx p(x)\Delta x.$$

$$\mathrm{Prob}(X_1+X_2) \in [x, x+\Delta x] = \int_{x_1=-\infty}^{\infty} \int_{x_2=x-x_1}^{x+\Delta x - x_1} p(x_1)p(x_2)\mathrm{d}x_2\mathrm{d}x_1.$$

As $\Delta x \to 0$ we obtain the convolution of p with itself:

$$\mathrm{Prob}(X_1+X_2 \in [a,b]) = \int_a^b (p*p)(z)\mathrm{d}z.$$

Exercise to show non-negative and integrates to 1.

Statement of Central Limit Theorem

- For simplicity, assume p has mean zero, variance one, finite third moment and is of sufficiently rapid decay so that all convolution integrals that arise converge: p an infinitely differentiable function satisfying
$$\int_{-\infty}^{\infty} xp(x)\mathrm{d}x \;=\; 0, \;\; \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x \;=\; 1, \;\; \int_{-\infty}^{\infty} |x|^3 p(x)\mathrm{d}x \;<\; \infty.$$
- Assume $X_1, X_2, \ldots$ are independent identically distributed random variables drawn from p.
- Define $S_N = \sum_{i=1}^{N} X_i$.
- Standard Gaussian (mean zero, variance one) is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Central Limit Theorem Let X_i, S_N be as above and assume the third moment of each X_i is finite. Then $\frac{S_N}{\sqrt{N}}$ converges in probability to the standard Gaussian:
$$\lim_{N\to\infty} \operatorname{Prob}\left(\frac{S_N}{\sqrt{N}} \in [a,b]\right) \;=\; \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\mathrm{d}x.$$

Proof of the Central Limit Theorem

- The Fourier transform of p is
$$\widehat{p}(y) \;=\; \int_{-\infty}^{\infty} p(x) e^{-2\pi i x y} \mathrm{d}x.$$
- Derivative of $\widehat{g}$ is the Fourier transform of $2\pi i x g(x)$; differentiation (hard) is converted to multiplication (easy).
$$\widehat{g}'(y) \;=\; \int_{-\infty}^{\infty} 2\pi i x \cdot g(x) e^{-2\pi i x y} \mathrm{d}x.$$
If g is a probability density, $\widehat{g}'(0) = 2\pi i \mathbb{E}[x]$ and $\widehat{g}''(0) = -4\pi^2 \mathbb{E}[x^2]$.
- Natural to use the Fourier transform to analyze probability distributions. The mean and variance are simple multiples of the derivatives of $\widehat{p}$ at zero: $\widehat{p}'(0) = 0$, $\widehat{p}''(0) = -4\pi^2$.
- We Taylor expand $\widehat{p}$ (need technical conditions on p):
$$\widehat{p}(y) \;=\; 1 + \frac{p''(0)}{2} y^2 + \cdots \;=\; 1 - 2\pi^2 y^2 + O(y^3).$$
Near the origin, the above shows $\widehat{p}$ looks like a concave down parabola.

Proof of the Central Limit Theorem (cont)

- $\mathrm{Prob}(X_1 + \cdots + X_N \in [a,b]) \ = \ \int_a^b (p * \cdots * p)(z)\mathrm{d}z$.
- The Fourier transform converts convolution to multiplication. If $\mathrm{FT}[f](y)$ denotes the Fourier transform of f evaluated at y:

$$\mathrm{FT}[p * \cdots * p](y) \ = \ \widehat{p}(y) \cdots \widehat{p}(y).$$

- Do not want the distribution of $X_1 + \cdots + X_N = x$, but rather $S_N = \frac{X_1 + \cdots + X_N}{\sqrt{N}} = x$.
- If $B(x) = A(cx)$ for some fixed $c \neq 0$, then $\widehat{B}(y) = \frac{1}{c}\widehat{A}\left(\frac{y}{c}\right)$.
- $\mathrm{Prob}\left(\frac{X_1 + \cdots + X_N}{\sqrt{N}} = x\right) \ = \ (\sqrt{N}p * \cdots * \sqrt{N}p)(x\sqrt{N})$.
- $\mathrm{FT}\left[(\sqrt{N}p * \cdots * \sqrt{N}p)(x\sqrt{N})\right](y) \ = \ \left[\widehat{p}\left(\frac{y}{\sqrt{N}}\right)\right]^N$.

Proof of the Central Limit Theorem (cont)

- Can find the Fourier transform of the distribution of S_N:

$$\left[\widehat{p}\left(\frac{y}{\sqrt{N}}\right)\right]^N.$$

- Take the limit as $N \to \infty$ for **fixed** y.
- Know $\widehat{p}(y) = 1 - 2\pi^2 y^2 + O(y^3)$. Thus study

$$\left[1 - \frac{2\pi^2 y^2}{N} + O\left(\frac{y^3}{N^{3/2}}\right)\right]^N.$$

- For any **fixed** y,

$$\lim_{N\to\infty} \left[1 - \frac{2\pi^2 y^2}{N} + O\left(\frac{y^3}{N^{3/2}}\right)\right]^N \;=\; e^{-2\pi y^2}.$$

- Fourier transform of $e^{-2\pi y^2}$ at x is $\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$.

Proof of the Central Limit Theorem (cont)

We have shown:

- the Fourier transform of the distribution of S_N converges to $e^{-2\pi y^2}$;
- the Fourier transform of $e^{-2\pi y^2}$ is $\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$.

Therefore the distribution of S_N equalling x converges to $\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$.

We need complex analysis to justify this conclusion. Must be careful: Consider

$$g(x) \;=\; \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

All the Taylor coefficients about $x = 0$ are zero, but the function is not identically zero in a neighborhood of $x = 0$.