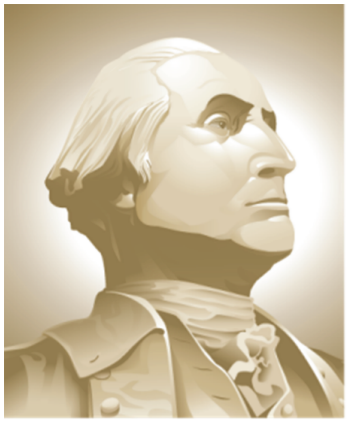
Lecture Notes EMSE 4765/6765: Statistical Analysis Review

Chapter 17: Basic statistical models

Version: 1/25/2021



THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Text Book: A Modern Introduction to Probability and Statistics, Understanding Why and How

By: F.M. Dekking. C. Kraaikamp, H.P.Lopuhaä and L.E. Meester

17 Basic Statistical Models

17.1 Introduction . . .

- **Most common statistical model:** Elements of the dataset are **repeated measurements of the same quantity** and **different measurements do not influence each other**.

- **Uncertainty in measurements: Modeled using a specific probability distribution (that we do not know)**. Random variables with this distribution represent the measurements prior to them being taken.

- **The measurements:** The actual observed measurement values are **called realizations of these random variables**.

- **Measurements not influencing each other:** A scenario where information of one measurement **does not provide any information** on the value of the next measurement. **Subsequent measurements are modeled as a random variables that are statistically independent, but with the same distribution.**

17 Basic Statistical Models

17.1 Random samples and statistical models . . .

Table 17.1. Michelson data on the speed of light.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

Source: E.N. Dorsey. The velocity of light. *Transactions of the American Philosophical Society.* 34(1):1-110, 1944; Table 22 on pages 60-61.

Data in Table 17.1 in $(Speed - 299,000)\ km/sec.$

Is there more than one "Speed of Light"? Answer: No, measurements reflect uncertainty of the measurement error

17 Basic Statistical Models

17.1 Random samples and statistical models . . .

- **Summarizing,** first measurement is modeled by a random variable X_1, second by a random variable X_2, etc. The value $x_1 = 850$ is interpreted as the realization of X_1 or the first datapoint, the value $x_2 = 740$ is interpreted as the realization of the random variable X_2 or the second data point, etc.

- **Moreover, one can think of X_1** as being uncertain about the value of measurement **prior to taking this measurement**. **One can think of x_1** as the value of the measurement **after it was taken.** Since experimental conditions are the same, one assumes that **the probability distributions** of X_1 and X_2 are **the same**. Since one measurement does not influence the other, X_1 and X_2 are assumed to be **statistically independent.**

Random Sample of size n: A collection of RV's $(X_1, X_2, \ldots, X_n)$ where $X_i \sim X$, with **some probability distribution** and X_i **are independent.**

17 Basic Statistical Models

17.1 Random samples and statistical models . . .

- **The random sample $(X_1, X_2, \ldots, X_n)$** can be thought of as **your plan to repeat an experiment n times of which the outcomes are uncertain.**

- **After executing this plan,** you have observed the dataset $(x_1, x_2, \ldots, x_n)$. One now calls the values $(x_1, x_2, \ldots, x_n)$ **a realization of the random sample $(X_1, X_2, \ldots, X_n)$.**

- **If you execute this plan over and over again,** each time you obtain **a different dataset** $(x_1, x_2, \ldots, x_n)$ with different sample mean, variance, etc.

- While **each random variable (planned measurement) possesses the same uncertainty model (the probability distribution of X), we do not know the specifics of this distribution.**

- **The field of Statistics** deals with, among many other things, **learning about this probability distribution of X,** through an observed dataset $(x_1, x_2, \ldots, x_n)$ from this **random sample $(X_1, X_2, \ldots, X_n)$, $X_i \sim X$.**

17 Basic Statistical Models

17.1 Random samples and statistical models . . .

- If $F(\cdot)$ is **the "parent" distribution function** of X in a random sample, we speak of *a random sample from* $F(\cdot)$. Similarly, we speak of a *random sample from a density* $f(\cdot)$, or *a random sample from a* $N(\mu, \sigma^2)$, etc.

Statistical model for repeated measurements: A dataset $(x_1, x_2, \ldots, x_n)$ of repeated measurements of the same quantity is modeled as **the realization of a random sample $(X_1, X_2, \ldots, X_n)$, $X_i \sim X$. The model may include a partial specification of the probability distribution of X.** This probability distribution is called **the model distribution.** Its parameters are called **the model parameters.**

Exercise : Suppose we obtain a dataset of ten elements by tossing a coin ten times and recording the result of each coin toss. **What would be an appropriate model distribution and what is the corresponding model parameter?**

Solution: We can say the realizations are of the values 0 (Heads) or 1 (Tails). Suppose then we have a dataset $(x_1, \ldots, x_{10}) = (1, 0, 0, 1, 1, 1, 0, 1, 0, 0)$. Defining next the random variables

$$X_i \equiv \text{The outcome of coin toss } i, \; i = 1, \ldots, 10$$

with each random variable $X_i \sim X$, with X posessing **the model distribution**:

$$Pr(X = x) = p^x (1-p)^{1-x}, \; x \in \{0, 1\} \Leftrightarrow Pr(X = x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases}$$

we say that, X is Bernoulli distributed with parameter p: $X \sim Ber(p)$.

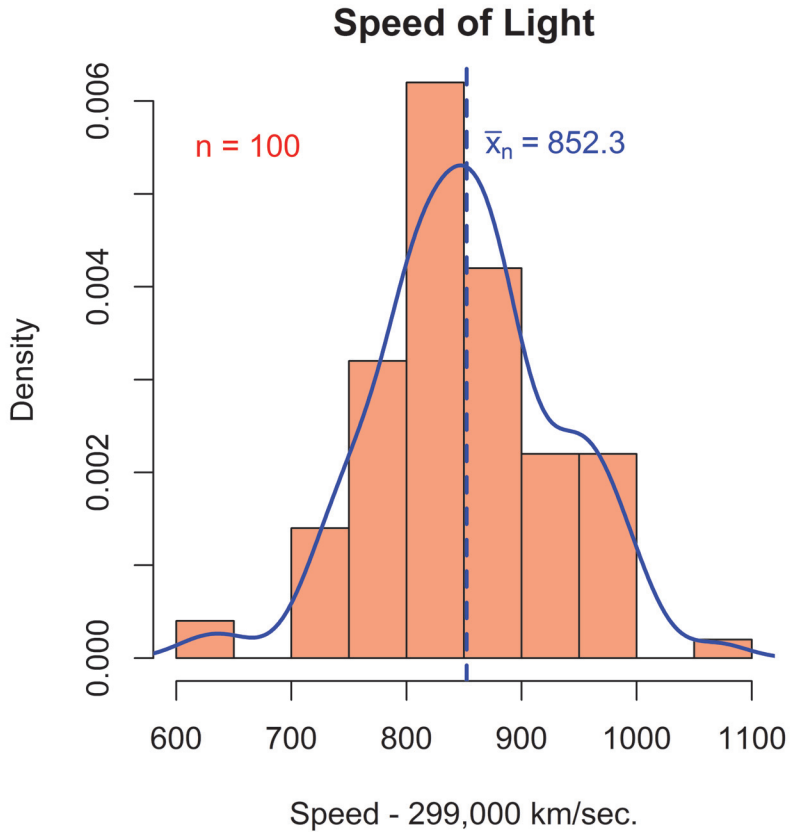
Each x_i coin toss outcome is thus a realization of X_i and the statistical model for the dataset $(x_1, \ldots, x_{10})$ is that it is **a realization of a random sample $(X_1, \ldots, X_{10})$**, where $X_i \sim Ber(p)$ and **the model parameter is p. One can learn about the value of the model parameter p** from **the observed dataset $(x_1, \ldots, x_{10})$.**

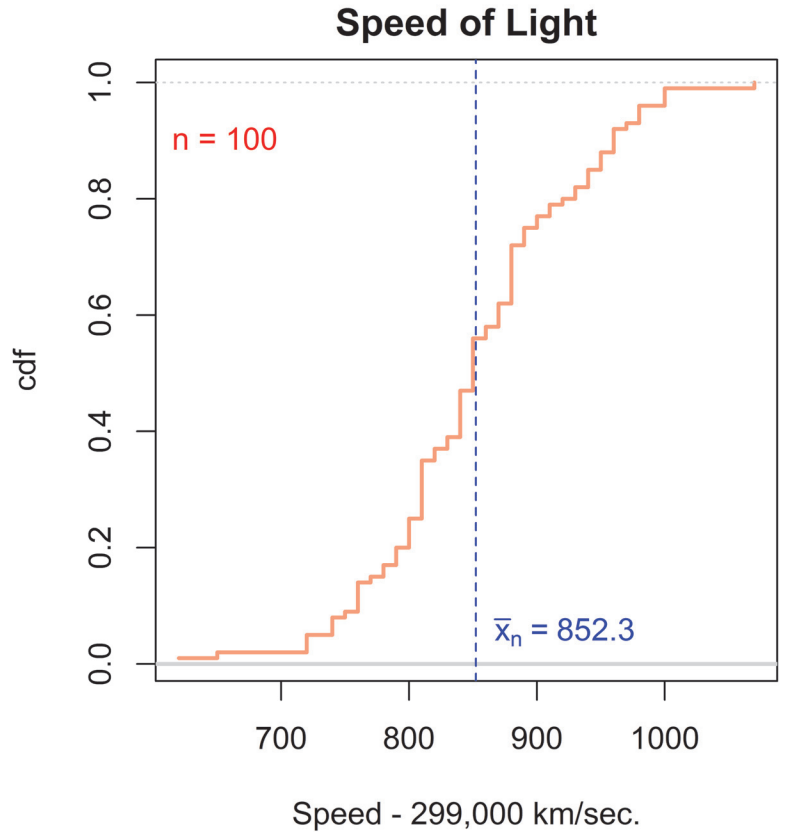
Use this basic statistical model to learn about unknown distributions through data:

Table 17.2. Some sample statistics and corresponding distribution features.

Sample statistic	Distribution feature
Graphical	
Empirical distribution function F_n	Distribution function F
Kernel density estimate $f_{n,h}$ and histogram	Probability density f
(Number of X_i equal to a)/n	Probability mass function $p(a)$
Numerical	
Sample mean $\bar{X}_n$	Expectation μ
Sample median $\text{Med}(X_1, X_2, \ldots, X_n)$	Median $q_{0.5} = F^{\text{inv}}(0.5)$
pth empirical quantile $q_n(p)$	100pth percentile $q_p = F^{\text{inv}}(p)$
Sample variance S_n^2	Variance σ^2
Sample standard deviation S_n	Standard deviation σ
$\text{MAD}(X_1, X_2, \ldots, X_n)$	$F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5)$, for symmetric F

Analysis in "Speed_of_Light.R"





- **Sample Statistic:** Since a dataset $(x_1, x_2, \ldots, x_n)$ of repeated measurements of the same quantity is modeled as **the realization of a random sample $(X_1, X_2, \ldots, X_n)$, the quantity $h(x_1, x_2, \ldots, x_n)$ is a realization or a data point from the object:**

$$T \equiv h(X_1, X_2, \ldots, X_n)$$

Thus T too is a Random Variable.
T is called a Random Sample Statistic or Estimator.

Note: For conciseness one ommits the word "random" and one refers to T as a sample statistic. But do not forget that T is in fact a random variable!

Sample Mean: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, **Sample Variance:** $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Empirical Distribution Function: $F_n(a) = \frac{\text{Number of } X_i\text{'s} \leq a}{n}$

- **Law of Large Numbers (LOLN):** Suppose $X \sim F(\cdot)$, $E[X] = \mu$, $V(X) = \sigma^2$

$$\lim_{n \to \infty} Pr(|\overline{X}_n - \mu| > \epsilon) = 0$$

$$\lim_{n \to \infty} Pr(|S_n^2 - \sigma^2| > \epsilon) = 0$$

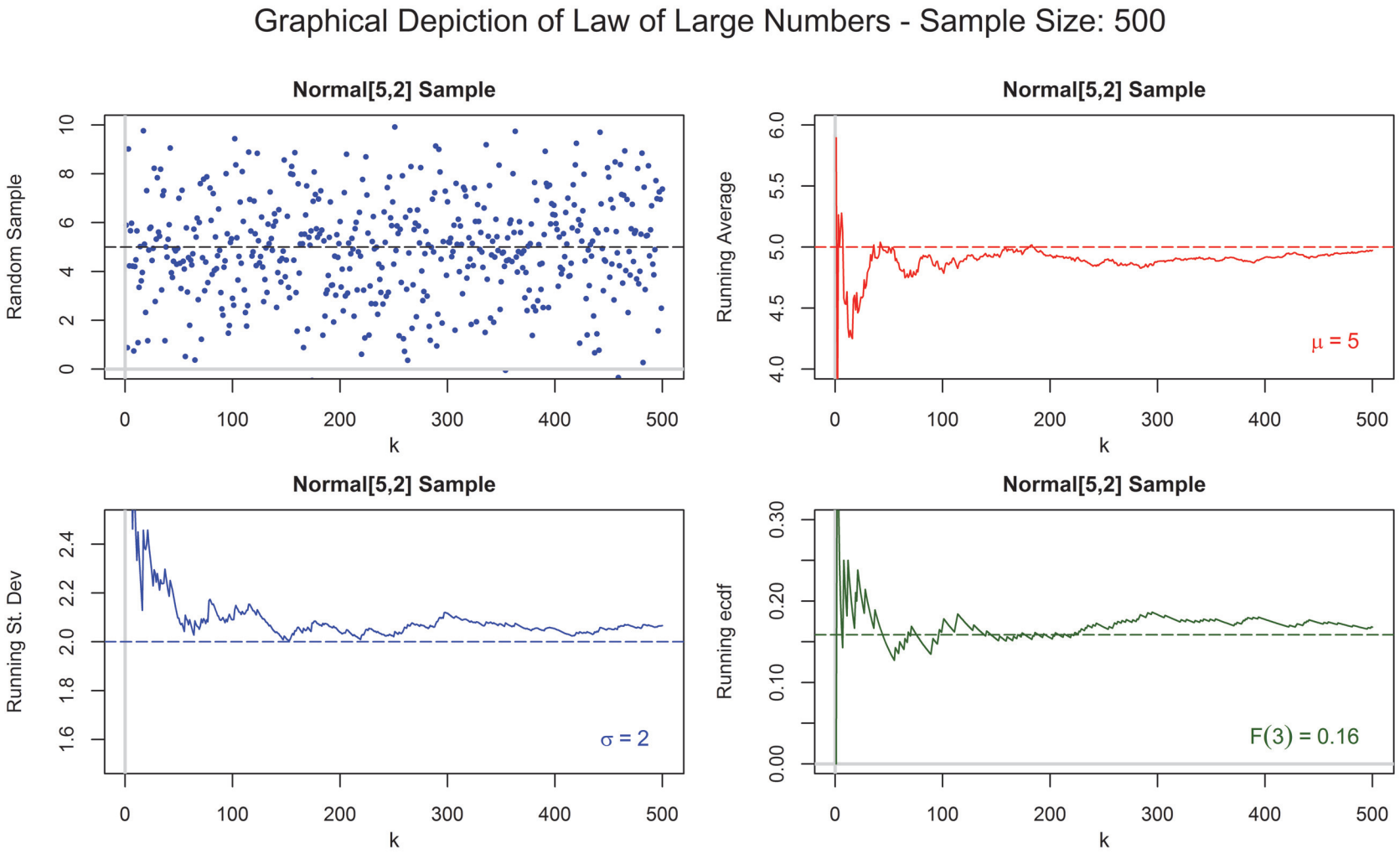
$$\lim_{n \to \infty} Pr(|F_n(a) - F(a)| > \epsilon) = 0$$

- **Conclusion:** Even though $\overline{X}_n$, S_n^2 and $F_n(a)$ are random variables, if the sample sizes are large, their realizations are close to the quantity of interest.

Example: Demonstration in *R*-file "Law_of_Large_Numbers_Normal.R" and Spreadsheet of Sample of size 500 from $N(5, 2)$.

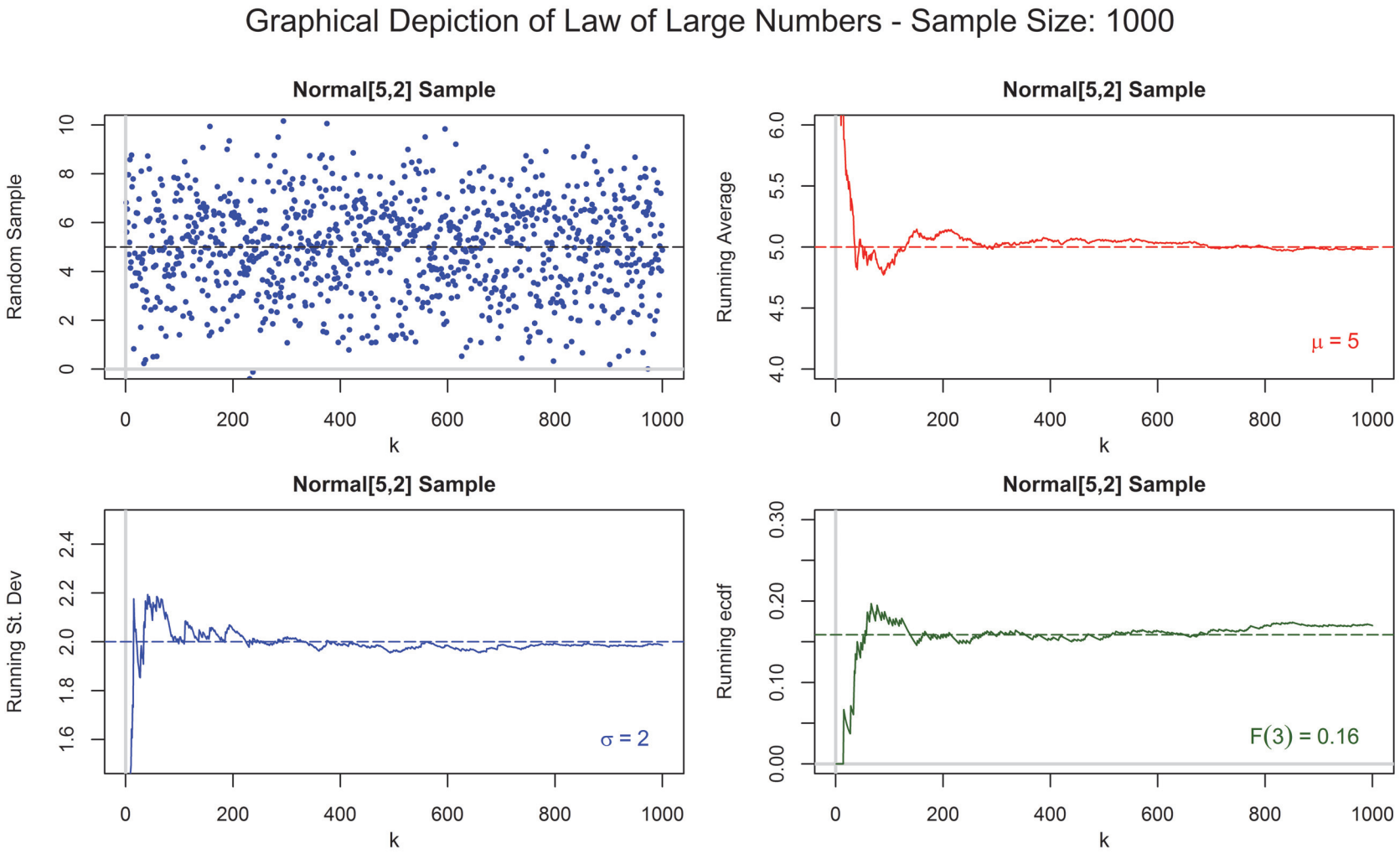
17 Basic Statistical Models

17.3 Estimating features of the "true" distribution



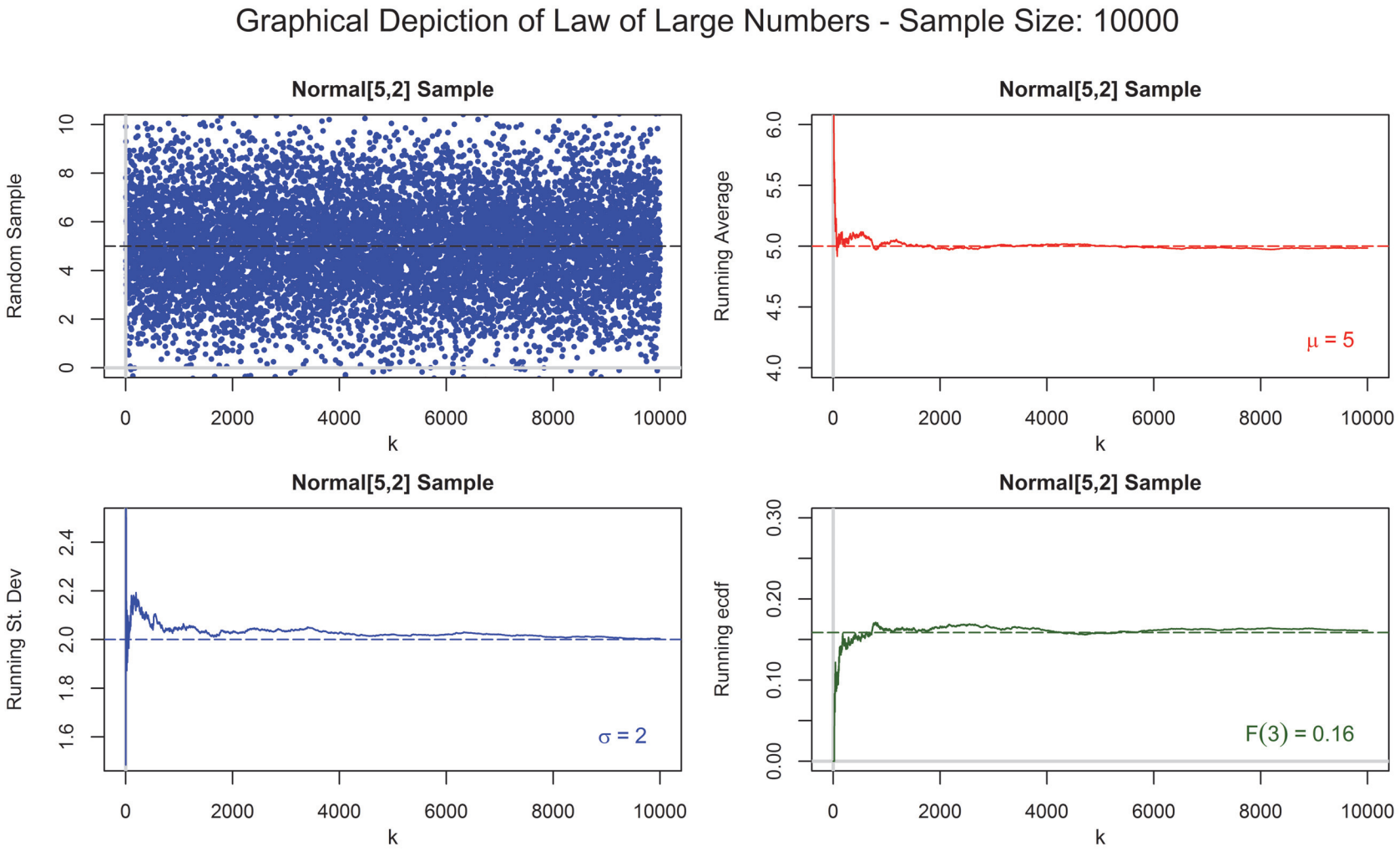
17 Basic Statistical Models

17.3 Estimating features of the "true" distribution



17 Basic Statistical Models

17.3 Estimating features of the "true" distribution



17 Basic Statistical Models

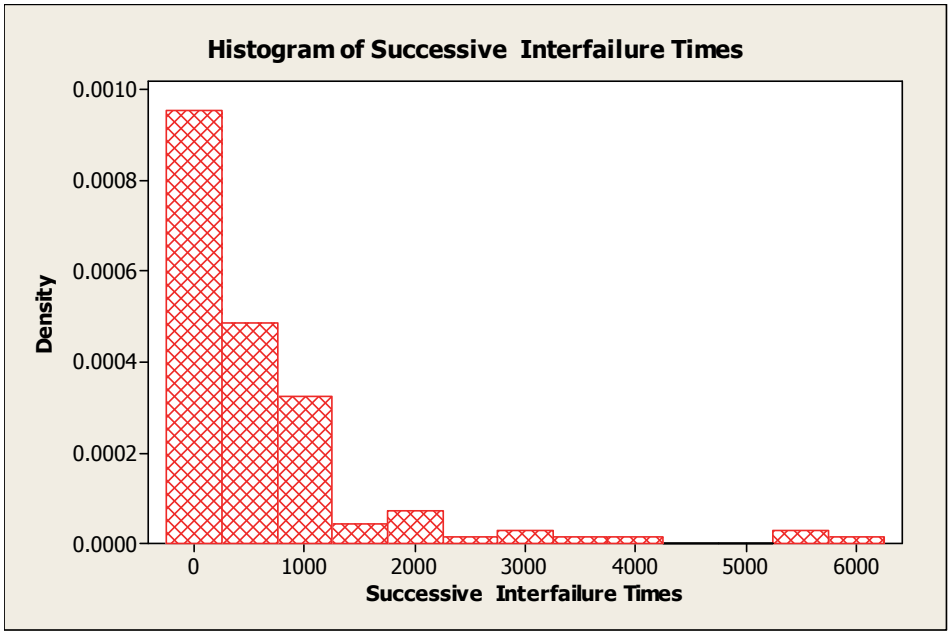
17.3 Estimating the "true" distribution

Example - Software Failure Times: In order to estimate the parameters of a software reliability model, failure data are collected. **The 136 failure times** are recorded, or equivalently, **the length of an interval between successive failures**. **Observed interfailure times in CPU seconds** for a certain software system.

Table 15.3. Interfailure times between successive failures.

30	113	81	115	9	2	91	112	15	138
50	77	24	108	88	670	120	26	114	325
55	242	68	422	180	10	1146	600	15	36
4	0	8	227	65	176	58	457	300	97
263	452	255	197	193	6	79	816	1351	148
21	233	134	357	193	236	31	369	748	0
232	330	365	1222	543	10	16	529	379	44
129	810	290	300	529	281	160	828	1011	445
296	1755	1064	1783	860	983	707	33	868	724
2323	2930	1461	843	12	261	1800	865	1435	30
143	108	0	3110	1247	943	700	875	245	729
1897	447	386	446	122	990	948	1082	22	75
482	5509	100	10	1071	371	790	6150	3321	1045
648	5485	1160	1864	4116					

Source: J.D. Musa, A. Iannino, and K. Okumoto. *Software reliability: measurement, prediction, application.* McGraw-Hill, New York, 1987; Table on page 305.



Example - Software Failure Times: Histogram suggest that failure time follows and exponential distribution: i.e. if we denote:

$$T \equiv \text{The software failure time } \in [0, \infty)$$

we have $T \sim Exp(\lambda)$ and for the model distribution

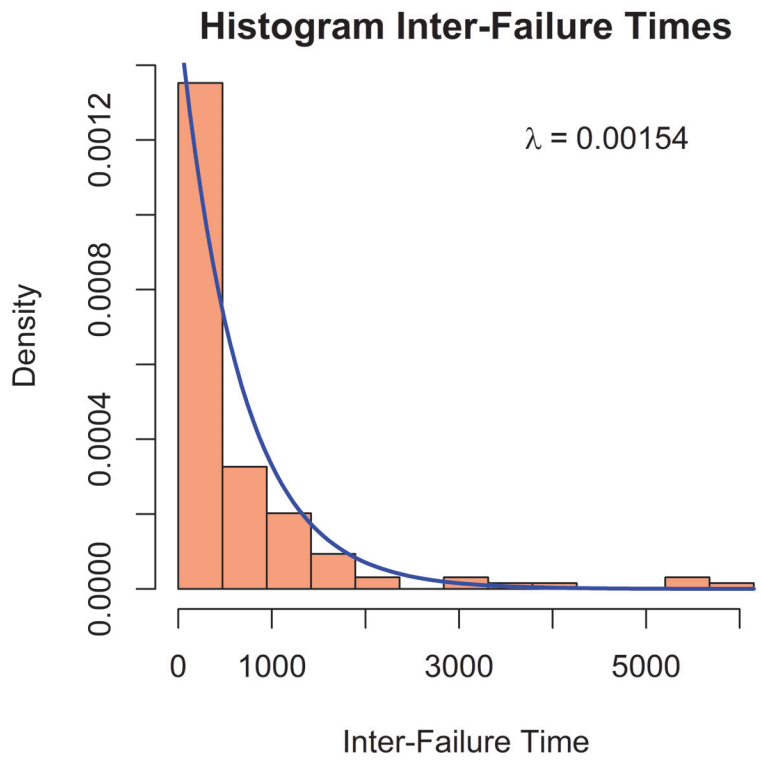
$$F(t) = P(T \leq t) = 1 - exp(-\lambda t), t \in [0, \infty), \lambda > 0.$$

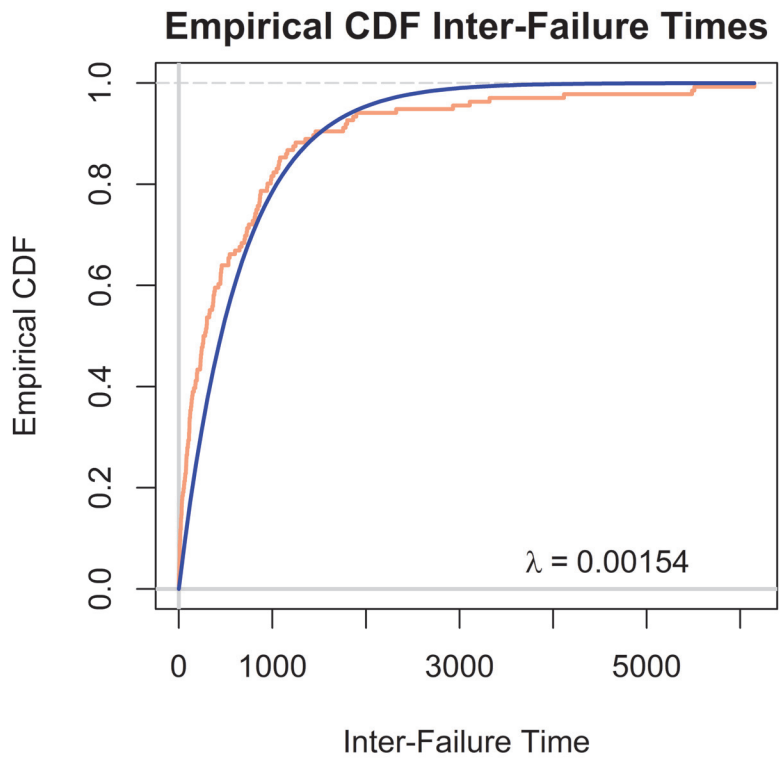
In this case we are interested in estimating the value of **the model parameter λ**, since given this value the distribution curve is completely specified.

Parametric Estimation Technique: We have:

$$\bar{t} = \frac{1}{136}\sum_{i=1}^{136} t_i \approx 649.14, \; E[T] = \frac{1}{\lambda} \Rightarrow \text{Set } \frac{1}{\lambda} = 649.14 \Rightarrow \lambda \approx 0.00154$$

Analysis in "SoftwareFailure_with_exp_fit"





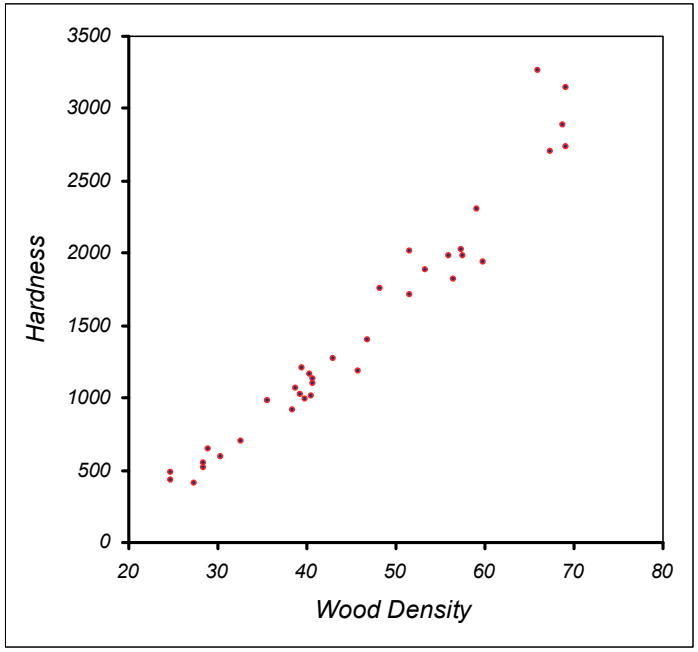
Example: Predicting hardness of Australian Timber

The Janka hardness test measures the hardness of wood. **To measure Janka hardness directly is difficult.** However, **it is related to the density of the wood**, which can be **measured without error**. In Table 15.5 **a bivariate dataset** is given of density (x) and Janka hardness (y) of 36 Australian eucalypt hardwoods.

Table 15.5. Density and hardness of Australian timber.

Density	Hardness	Density	Hardness	Density	Hardness
24.7	484	39.4	1210	53.4	1880
24.8	427	39.9	989	56.0	1980
27.3	413	40.3	1160	56.5	1820
28.4	517	40.6	1010	57.3	2020
28.4	549	40.7	1100	57.6	1980
29.0	648	40.7	1130	59.2	2310
30.3	587	42.9	1270	59.8	1940
32.7	704	45.8	1180	66.0	3260
35.6	979	46.9	1400	67.4	2700
38.5	914	48.2	1760	68.8	2890
38.8	1070	51.5	1710	69.1	2740
39.3	1020	51.5	2010	69.1	3140

Source: E.J. Williams. *Regression analysis.* John Wiley & Sons Inc., New York, 1959; Table 3.1 on page 43.



- The idea is, of course, that Janka hardness is related to the density: the higher the density of the wood, the higher the value of Janka hardness.

$$y \equiv hardness,\ x \equiv density \Rightarrow y = g(x)$$

- **Scatter plots suggests :**

$$g(x) = \alpha + \beta \cdot x$$

- Scatter plots also suggests the linear relationship **is not perfect** and **contains measurement error.**

- **Regression Model:** Let U be a random error term, then we can model

$$Y = \alpha + \beta \cdot x + U \Leftrightarrow U = Y - \alpha + \beta \cdot x$$

- The line $y = \alpha + \beta \cdot x$ is called **the regression line**, x is called **the explanatory variable** and Y is called **the (random) dependent variable.**

Simple Linear Regression Model: In a simple linear regression model for a bivariate dataset $(x_1, y_1), \ldots, (x_n, y_n)$, we assume that $x_1, \ldots, x_n$ are **non-random** and that $y_1, \ldots, y_n$ are **realizations** of **random variables $Y_1, \ldots, Y_n$** satisfying

$$Y_i = \alpha + \beta x_i + U_i \Rightarrow E[Y_i] = \alpha + \beta x_i + E[U_i],\ V[Y_i] = V[U_i]$$

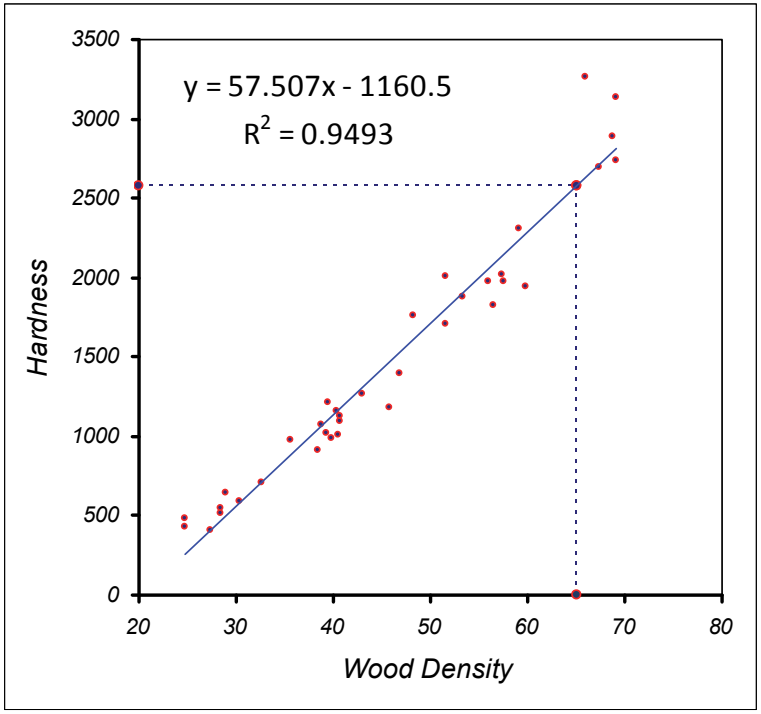
for $i = 1, \ldots, n$ where $U_i, i = 1, \ldots, n$ are independent RV's and

$$U_i \sim U,\ E[U] = 0,\ V(U) = \sigma^2,\ u_i = y_i - \alpha + \beta x_i,\ i = 1, \ldots, n$$

Exercise:
Is $(x_1, \ldots, x_n)$ a realization of a random sample $X_1, \ldots, X_n$? Is $(u_1, \ldots, u_n)$ a realization of random sample? Is $(y_1, \ldots, y_n)$ a realization of a random sample?

Exercise: Suppose we have a eucalypt hardwood tree with density 65. What would your prediction be for the corresponding Janka hardness?

Answer:



Step 1: Add **linear trendline** in MicroSoft Excel to find equation estimate:

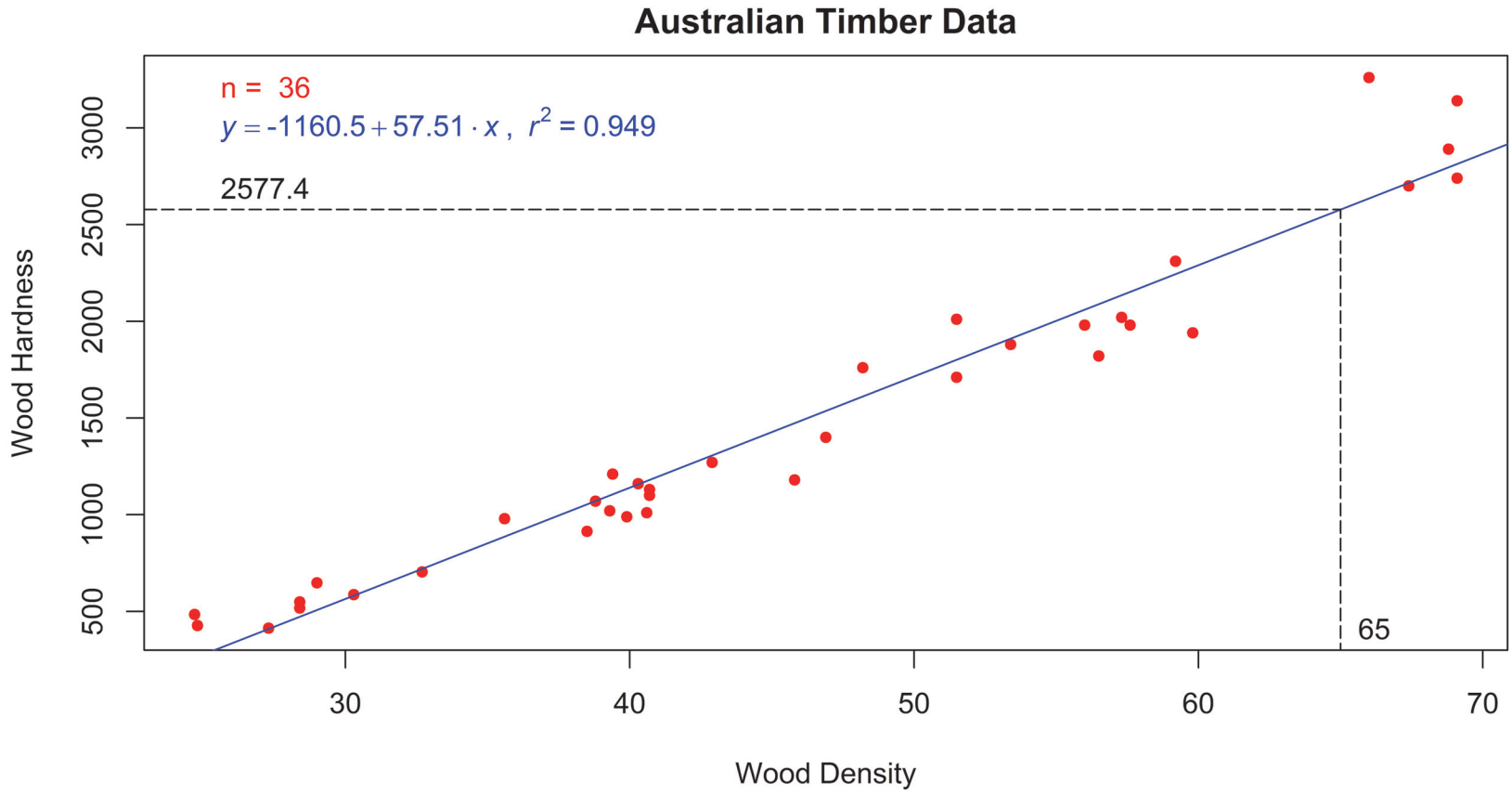
$$y = 57.707x - 1160.5$$

Step 2:

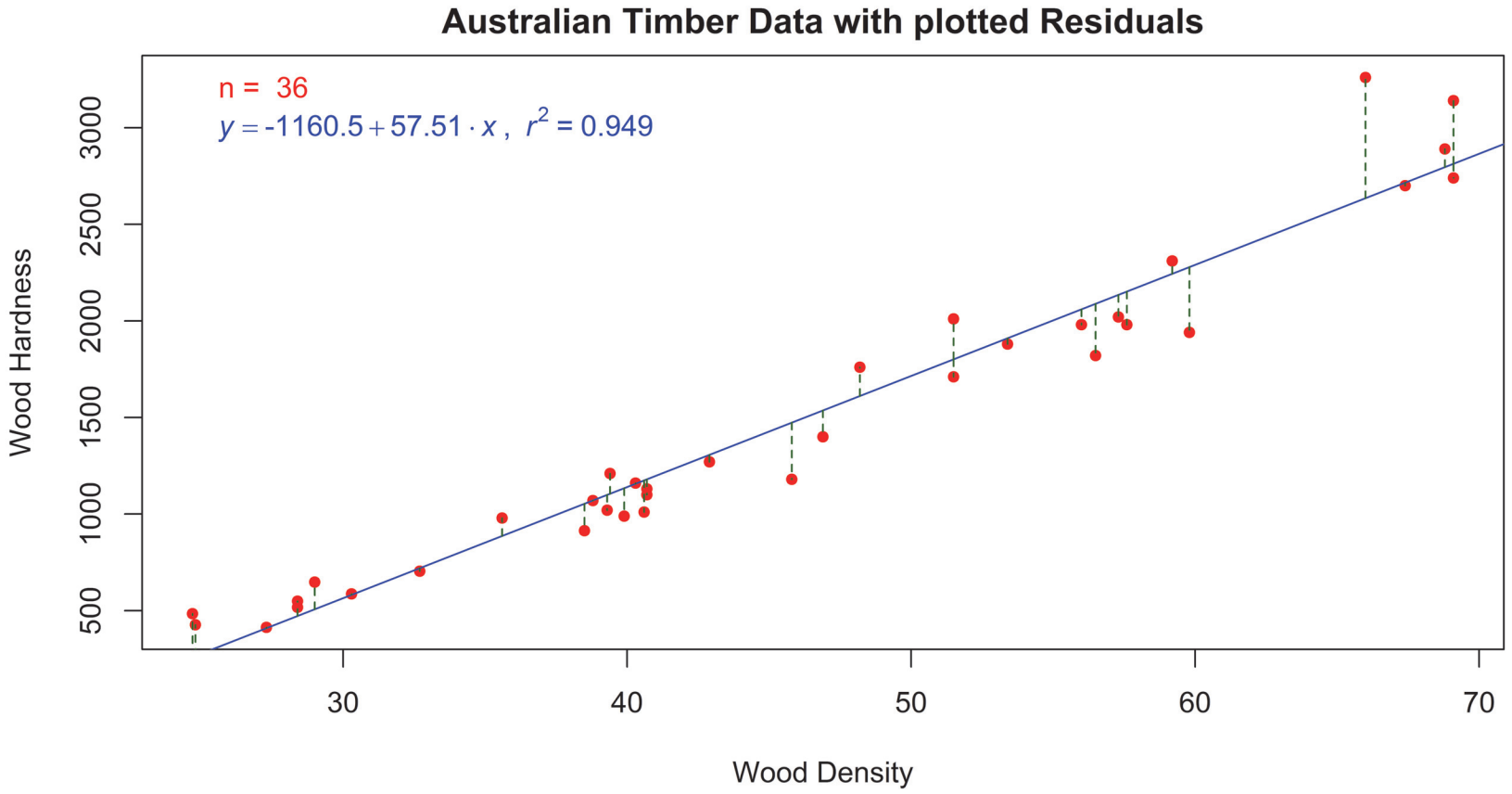
$$x = 65 \Rightarrow y \approx 2577$$

What about the uncertainty in $Y(65)$?

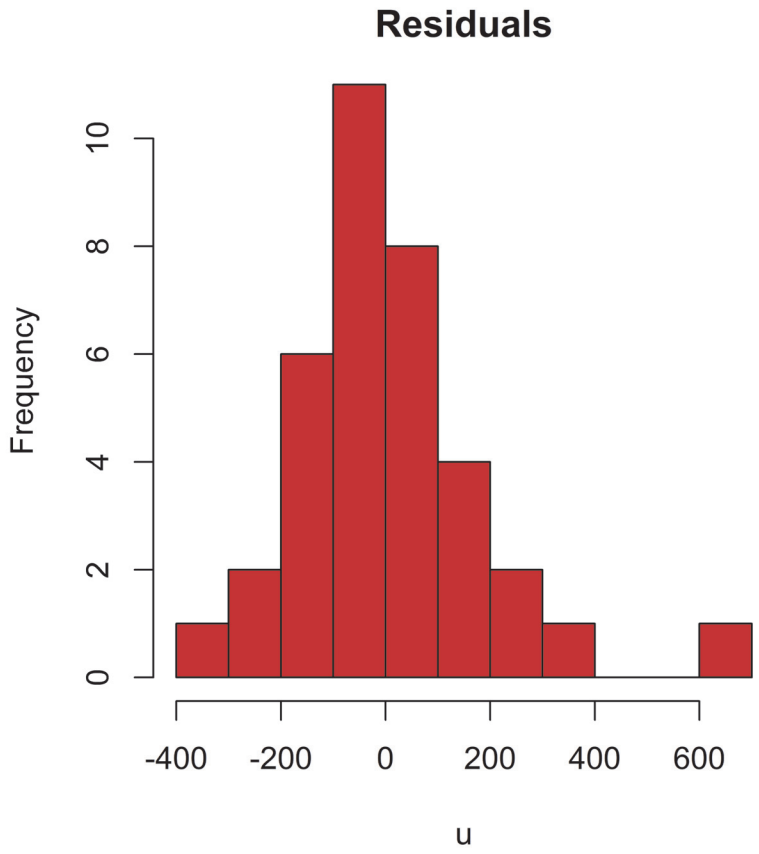
Same Analysis in *R*-file "AustralianTimber_Prediction.R"

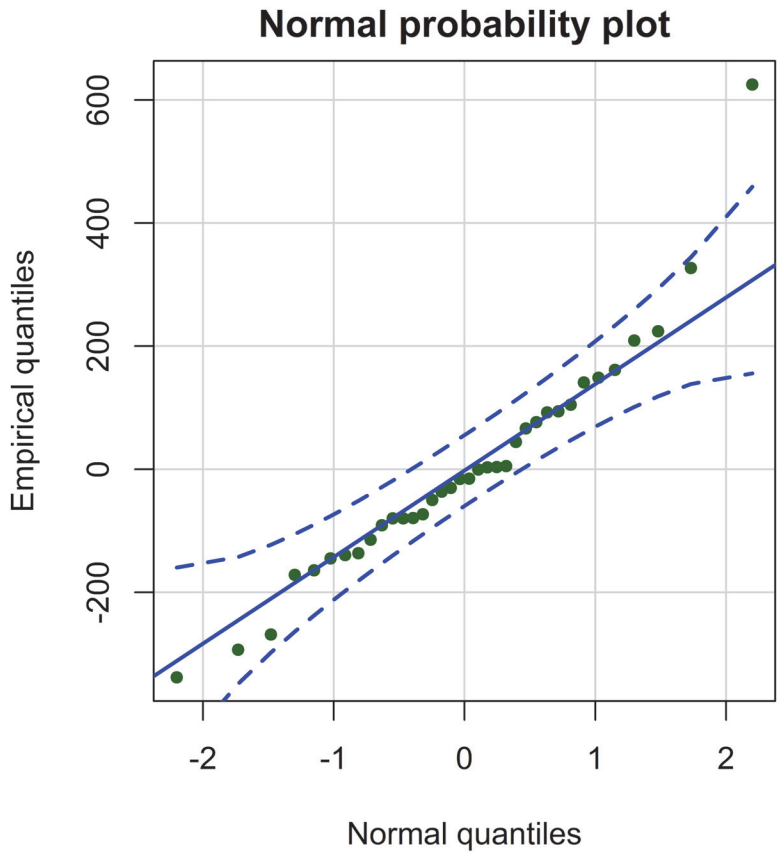


Analysis of $(u_1, \ldots, u_n)$ in R-file "AustralianTimber_Prediction.R"



Histogram of $(u_1, \ldots, u_n)$ for U in R-file "AustralianTimber_Prediction.R"





Analysis Prediction Uncertainty $Y(65)$ in R-file "AustralianTimber_Prediction.R"