2009

Glacial Lake Missoula And The Self Censoring Data Problem

Cody L. Custis

Submitted in partial requirements of a Master Of Science Degree in Statistics at Montana State University

Version 3.3

Creative Commons

This page is available in the following languages:
български Català Dansk Deutsch English English (CA) English (GB) Castellano Castellano (AR) Español (CL) Castellano (MX) Euskara Suomeksi français français (CA) Galego עברי hrvatski Magyar Italiano 日本語 한국어 Melayu Nederlands polski Português svenska slovenski jezik 简体中文 華語 (台灣)

Attribution-NonCommercial-ShareAlike 2.5

You are free:

to copy, distribute, display, and perform the work
to make derivative works
Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.
For any reuse or distribution, you must make clear to others the license terms of this work.
Any of these conditions can be waived if you get permission from the copyright holder.
Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

Learn how to distribute your work using this license

http://creativecommons.org/licenses/by-nc-sa/2.5/

Abstract

A self-censoring transformation is one in which future observations censor current observations if the future observations are larger than the current ones. As a result, it is possible that most of the original data is lost after the transformation is applied. One prominent example that suggests self-censoring is waterlines left from Glacial Lake Missoula on the hillsides which surround Missoula, Montana.

One of the major controversies concerning Glacial Lake Missoula is the number of floods. Previous research has generally ignored the waterlines, due to difficulty of estimating the number of original observations given a self-censored count. Simulation is used to determine the count of self-censored waterlines based on different numbers (and methods) used to generate the non-censored heights of the waterlines.

Counts are compared between temporally dependent processes and non-temporally dependent processes. It is found that the counts are much higher, and one can differentiate between dependent and independent processes using logistic regression to predict the process of the original observations based on the count after self-censoring is applied.

The relationship between the count after self-censoring and the original number of observations is explored for data generated from a random walk distribution. Regression is used in an attempt to relate the number of original observations to the count after self-censoring is applied.

The Kolmogorov-Smirnov test is used to attempt to differentiate between empirical cumulative distribution functions (e.c.d.f.s) after self-censoring and standardization transformations are applied. However, the test is found to perform poorly, mainly due to the small counts after self-censoring is applied.

Outline

Title Page

Copyright

Abstract

Outline

Count.Orig.R.txt
K-S.R.txt

Appendix 3: Selected Graphs
(Numbers in lower right)
III.2.1 Random Walks
III.5.1 Empirical CDF From Exponential Distribution
III.5.2 Empirical CDF From Normal Distribution
III.5.3 Empirical CDF From Uniform Distribution
III.5.4 Empirical CDF From Random Walk Distribution
III.6.1 Boxplots Comparing Counts After Self-Censoring
III.6.2 Logistic Regression Curve
III.6.3 Histogram of Counts
III.7.1 Counts As A Function Of Original Observations
III.7.2 Counts As A Function Of Original Observations, Expanded
III.9.1 Example c.d.f.s For K-S Test
III.9.2 Probability Of Rejection As a Function of Significance, Exponential vs. Uniform
III.9.3 Probability Of Rejection As a Function of Significance, Exponential vs. Exponential
III.9.4 Probability Of Rejection As a Function of Significance, Exponential vs. Walk
III.9.5 Probability Of Rejection As a Function of Significance, Exponential vs. Walk (Fewer Observations)
III.10.1 Probability Of Rejection As a Function of Significance, Normal vs. Normal, Self-Censoring
III.10.2 Probability Of Rejection As a Function of Significance, Walk vs. Walk, Self-Censoring
III.10.3 Probability Of Rejection As a Function of Significance, Exponential vs. Walk, Self-Censoring
IV.1.1 Counts From Waittian Hypothesis
IV.1.2 Counts From Atwaterian Hypothesis

Appendix 4: Bibliography

I. Introduction

1. Glacial Lake Missoula Introduction

Having spent some time in Missoula, I am quite aware of the hills that surround the city. Mount Sentinel is part of the only college campus in the nation with a mountain on campus. Mount Jumbo is owned by the city of Missoula, the result of Missoula's efforts to save open space. At first glance, most will notice the letters on the hills, the "M" on Mount Sentinel and the "L" on Mount Jumbo. Perhaps one notices the hiking trails, usually covered in Missoulians looking to get exercise. Perhaps one notices the trees that top Mount Sentinel, which have grown down due to years of fire suppression. And maybe, if the weather is right, one notices horizontal stripes going across both Mount Sentinel and Mount Jumbo. One wonders what caused these stripes.

Imagine yourself a member of one of the native tribes that inhabited the Northwest during the last Ice Age. On a relatively warm summer's day, you take your canoe out to the lake near the village where you live, perhaps on your way to hunt game. All of a sudden, the level of the lake drops dramatically. You paddle furiously against powerful currents, fighting to make it to the lakeshore. Taking refuge in one of the trees that border the lake, you watch as the lake which you had always known from a thousand foot deep lake to a mere river in a period of days. You see the ice dam, which had held the water, burst open, releasing a torrent of water on the landscape, along with other debris from the lake.

The lake which you had known, Glacial Lake Missoula, once extended from near Sandpoint Idaho, up to Missoula and the five valleys that converge in the Missoula valley. Evidence for Glacial Lake Missoula comes from a variety of sources. There is evidence from layers deposited by the massive floods, which can be found downstream in Eastern Washington. Some have investigated the layers of sediment left by the yearly cycles of the lake at Ninemile Creek, a mere twenty miles from Missoula. Glacial Lake Missoula also left evidence in those horizontal stripes, waterlines formed by erosion. It is those waterlines that are perhaps the most visible evidence of Glacial Lake Missoula, visible to all who inhabit the Missoula valley in present day.

While the idea of Glacial Lake Missoula was once so outrageous that the geological community shunned Bretz, one of the major geologists who researched Glacial Lake Missoula, it is now well accepted. However, knowing that Glacial Lake Missoula happened does not answer a very important question, how many times did Glacial Lake Missoula happen? This is a more difficult question. Evidence comes from sediment layers. However, the evidence from sediment layers appears to vary depending on which site the data is collected from. Likewise, other floods occurred in the area, so perhaps the depositional layers do not come from the Glacial Lake Missoula flood but from other floods. Likewise, the waterlines on the hillside of the Missoula valley are known to come from Glacial Lake Missoula. However, waterlines are notoriously easy to erase, unable to survive submersion in water.

When J. Harlen Bretz first proposed Glacial Lake Missoula, he offered a theory that few floods occurred. Recent geologists such as R.B. Waitt have proposed many floods. J. Shaw has proposed a theory that only a few floods occurred, perhaps a single flood. The goal of this paper is to examine the different geological theories, and use those theories to generate the pattern of waterlines. Geological theory provides the mechanisms by which waterlines would be generated, but statistical techniques will allow examination of self-censored data through the use of simulation.

2. Description Of Glacial Lake Missoula

Glacial Lake Missoula was a lake formed by the damming of the Clark Fork river by the Purcell Trench lobe of the Cordilleran ice sheet. With no other outlets, the water rose to a depth of thousands of feet, filling the valleys of western Montana. Based on the shorelines, geologists calculated that

glacial Lake Missoula covered roughly 2,900 square miles of Western Montana, with a depth of over 2,000 feet and volume of roughly 500 cubic miles.

Based on ash from Mount Saint Helens found in sediment layers from Glacial Lake Missoula, it is known that the floods occurred roughly 15,000 years ago. Based on varves, formed by sediment on the bottom of lakes, Alt and Chambers believe that Glacial Lake Missoula lasted for roughly 1,000 years. [in Alt, 2002]

Ice is a problematic substance to build a dam. Based on physics and observations of modern ice dams, ice dams fail catastrophically when the height of the water is roughly .9 the height of the ice dam. [Waitt, 1985] This is due to the pressure of the water forcing tiny cracks in the ice. As water flows through the ice, it generates heat, which leads to further melting of the ice. Thus, ice dams fail catastrophically, starting at the bottom. After the ice dam fails, most of the volume of the lake is released.

Modern lakes held by ice dams are relatively small. However, during the last ice age, much larger lakes formed. Glacial Lake Missoula is among the largest of such lakes, and perhaps the most controversial, due to the linking of Glacial Lake Missoula with the channeled scablands of Eastern Washington.

Glacial Lake Missoula drained very quickly, probably in period of three days of less. [Alt, 2002] The massive volume of water rushed over Eastern Washington, cutting the landscape, depositing sediments, leaving boulders in its path, and leaving giant ripples in Camas prairie. The water rushed over the landscape until the Wallula Gap (near Pasco, Washington), where only 40 cubic miles of water per day could be discharged. The volume of water was once again held up at the Kalama narrows, resulting in flooding of the Willamette valley in Oregon from Glacial Lake Missoula.

A more recent controversy concerns the number of such floods. Bretz initially proposed a single flood to explain the channeled scablands. However, Bretz increased his estimates over time. After Bretz's work, other geologists started to propose theories of fewer floods. However, in the 1970s and early 1980's, geologists would drastically increase the estimates of the number of floods. Most geologists agree that there were many floods, but those estimates range from 40 up to 80.

3. Pardee's Hypothesis

The vast majority of the research on the Glacial Lake Missoula floods has focused on sediment layers deposited downstream. Yet, the original inspiration for Glacial Lake Missoula was not found from sediments downstream, but on the hillsides of the Missoula valley.

Those that visited the Missoula valley were long aware of such lines, which appeared to come from some sort of body of water. These lines extend through out the Missoula valley, down the Bitterroot valley, and even up to the Jocko valley. [Pardee, 1910].

As far as the substance of such lines, "Douglass describes these benches as beds of sand, gravel, and volcanic ash of Miocene age in part. Remnants of the 'trails' are not only preserved upon open slopes of this easily eroded material, but upon the sides of ravines that dissect it, showing that time since the 'trails' were formed has been too brief for any material alteration of the topography by erosion." [Pardee, 1910]

J.T. Pardee is credited with explaining this phenomenon as due to a prehistoric lake. "The foregoing phenomena as a whole seem explainable only as the records of an extinct lake or sea. The old 'buffalo trails' are the existing remnants of its wave terraces. Its high level was approximately 4,200 feet above sea. At this stage the site of the present city of Missoula was 1,000 feet under water, and glaciers from the Bitter Root range south of Hamilton reached the lake, setting bowlder-laden icebergs afloat upon it." [Pardee, 1910]

Likewise, Pardee recognized that such a lake could be formed by ice. "The evidence of icebergs, together with the apparent recency of the lake and the variable height of its surface, connect this lake with the glacial period, and readily lend themselves to the suggestions that its dam was of ice." [Pardee, 1910]

Thus, Pardee effectively linked the waterlines on the hillsides with a lake formed during the most recent ice age.

II. Geological Background

1. The Views Of J. Harlen Bretz

While Pardee is given credit for popularizing Glacial Lake Missoula, due to his research on the waterlines, J. Harlen Bretz perhaps did the most to bring fame to Glacial Lake Missoula. Bretz linked the "channeled scabland" of Eastern Washington with massive bodies of water, and offered the controversial explanation that the channeled scabland was caused by events that occurred relatively few times. This went against the current view in geology that landforms were the result of processes that were repeated over history, uniformity. Eventually, the geological community came around to Bretz's position.

Bretz examined the unique geographical features of Eastern Washington. After numerous trips to the region, Bretz needed an explanation for the landforms found in Eastern Washington.

"(The scablands of Eastern Washington) have a history which is believed to be unique. The prevailing feature of their topography is indicated in the term used here: channeled scablands."

"(The) unique combination of topographical features of the Columbia Plateau in Washington has only one interpretation consistent (with twenty features that Bretz describes). *The channeled scablands are the erosive record of large, high-gradient, glacier-born streams.*" (Italics added) [Bretz, 1924]

Bretz came to the conclusion that the channeled scablands of Eastern Washington were due to some form of massive flooding.

"Furthermore, gradients were high and the glacial waters eroded enormously, sweeping away the overlying loessial material, crossing low divided and isolating many groups of the maturely eroded hills to produce the anastomosing pattern of the scablands, biting deeply into the basalt to make the canyons and rock basins, and spilling into the Snake and Columbia in three times as many places as the pre-existing drainage has used...." [Bretz, 1924]

"From one of these canyons alone 10 cubic miles of basalt was eroded by its glacial stream." [Bretz, 1924]

Bretz proposed that the landforms of Eastern Washington were due to massive floods. Unfortunately, Bretz did not have a source for this water. Indeed, many suggested (with a fair amount of derision), that Bretz was searching for evidence of the great flood of Genesis. Both Bretz and Pardee had considered proposing Glacial Lake Missoula as a source for the "Spokane floods." However, another geologist, HT Harding, would be the first to publish a hypothesis that linked the channeled scabland of Eastern Washington with Glacial Lake Missoula.

"A means of providing this volume of water, approximately fifty cubic miles per day, has puzzled Dr. Bretz and other geologists interested in the phenomenon. Dr. Bretz gives two possible explanations of such a flood: first, that the glacier was melted excessively in a short period of time by extreme heat and warm rains; or second, that the Spokane flood was a gigantic Jokulloup, such as has occurred in Iceland where volcanic activity has broken out beneath and ice-cap." [Harding, 1929]

However, Harding proposed an alternative source for such water.

"I am more inclined to believe that the flood was caused in a more natural and regular way, by the ponding of waters in natural reservoirs by ice obstructions.

The essential factors necessary to supply a sufficient amount of water to cause the channeled scabland are as follows: first, a secure ice dam in the Columbia River below the Nesplem River to divert the water across the eastern Washington table-land; second, a reservoir large enough to hold a sufficient amount of water; third, a temporary dam sufficiently high to impound the water; fourth, water to fill the reservoir." [Harding, 1929]

Harding goes on to suggest Glacial Lake Missoula as a possible source for the water. Importantly for our question, Harding suggested that Glacial Lake Missoula was an event that occurred

in regular cycles. "It is the belief of the writer that during the different glaciation periods many floods occurred which were produced in a similar manner (from the breaking of glacial lakes)." [Harding, 1929] From a Hardinian perspective, the number of floods would be limited only by the length of time in which the proper conditions for catastrophic floods held. Thus, one could base an estimate on information about the volume of the glacial lake, when temperatures were sufficiently cold for the ice dam to form, and how long it took to fill such a lake.

Bretz would accept Glacial Lake Missoula as the source of the water needed to form the channeled scablands of Eastern Washington. However, Bretz believed that there was a single great Glacial Lake Missoula flood, due to glacial retreat.

"It is suggested that bursting of the ice barrier which confined a large glacial lake among the mountains of western Montana suddenly released a very great quantity of water which escaped across the plateau of eastern Washington and eroded the channeled scablands.

Lake Missoula, named and first described by J.T. Pardee, occupied a large number of connected intermontane valleys in Montana between the continental divide and the panhandle of Idaho. It rose to 4,200 feet A. T. (altitude) and had a maximum depth of at least 2,100 feet. At the extreme northwestern end of the dam the ice was crowded against the steep north-facing terminus of the Bitterroots in the southern bifurcation of Purcell Trench. A retreat of two miles here could empty the entire lake if there was free escape afterward." [Bretz, 1930]

While both Harding and Bretz linked Glacial Lake Missoula to the channeled scablands, there remains one important difference in the views of Harding and Bretz. Bretz proposed that Glacial Lake Missoula burst due to retreat of glaciers, suggesting that there was a single Glacial Lake Missoula flood. Harding proposed a theory where Glacial Lake Missoula burst due to failure of the ice dam, and would reform after each flood, meaning that floods were periodic.

No discussion of Glacial Lake Missoula would be complete without a discussion of the ensuing controversy involving Bretz. Bretz would accept the hypothesis of Harding and do additional field work to link Glacial Lake Missoula to the channeled scablands of Eastern Washington. Much of this was done through additional field work, showing that the geography of Eastern Washington was due to flooding. [Baker, 1978] Nonetheless, the geological community refused to accept Bretz's theory that the landscape was due to a few large floods.

However, in 1940, J.T. Pardee would settle the controversy.

"Pardee's written paper also had an understated title: 'Unusual Currents in Lake Missoula.' His work, which began before Bretz's studies, clearly demonstrated that Lake Missoula was the source of catastrophic floods though the Channeled Scabland. He noted that about 2000 km^3 of water was impounded behind a glacial lobe that occupied the basin of modern Pend Oreille Lake in northern Idaho. Pardee believed that this glacial dam had failed suddenly, with a resultant rapid draining of the lake. Evidence of this failure included severely scoured constrictions in the lake basin, huge bars of current transported debris, and the giant current ripple marks." [Baker, 1978]

Further field work by Bretz in the 1950s would provided additional evidence confirming the hypothesis that the channeled scablands were caused by great floods, and evidence against alternative hypotheses. Bretz was long lived, and saw his once controversial hypothesis accepted by the geological community nearly forty years after it was first proposed. In 1979, at 96 years of age, Bretz received the Penrose Medal, the Geological Society of America's highest award. He is quoted as saying "all my enemies are dead, so I have no one to gloat over."

Nonetheless, this leaves one very important question. Although the geologic community agreed that Glacial Lake Missoula was responsible for the channeled scabland of Eastern Washington, that does not answer the important question of how many times Glacial Lake Missoula occurred. However, one legacy of J Harlen Bretz is that most of the work on such a question has involved field sites in Eastern Washington. Initially, most geologists agreed with Bretz that the number of Glacial Lake Missoula floods was small. It would not be until the 1980s that geologists returned to the Harding view of many floods.

2. The View From The 1950s Onward

Most of the research into Glacial Lake Missoula from the 1950s to 1980 assumed that there were a very small number of floods. Different numbers of floods were needed to explain certain features. However, the general trend was to use few floods to explain the landscape.

"A lower limiting radiocarbon date on scabland flooding was established by Fryxell (1962). Near Vantage, Washington, wood dated at 32,700 + / - years B.P. was found in deposits of the major flood.... If the correlation to the early Pinedale of Richmond's (1965) Rocky Mountain glacial chronology is valid, a reasonable data for *the* flood would be about 22,000 years B.P.

Pardee (1942) noted that Lake Missoula formed at least once after the catastrophic drainage, achieving only one-third of the former lake volume.... Thus, a situation opportune for a catastrophic breakout flood was present during middle Pinedale time (about 14,000 years B.P.). Late Pinedale ice (9,000 years B.P.) did not advance far enough south to form a glacial Lake Missoula.

Bretz, Smith, and Neff (1956) and Bretz (1969) have described several features of the Columbia River Valley, west of the Columbia Plateau, which suggest that an episode of flooding occurred after the early Pinedale flood. Possibly this flood encountered no Okanogan ice lobe, which would have diverted it across the Columbia Plateau. It would, therefore, have been confined solely to the valley of the Columbia River. Further study is required to determine whether this flood is related to the middle Pinedale formation of Lake Missoula." (Italics mine) [p. 12 Baker 1973]

Indeed, the common view from the 1950s until 1980 was that there was only a single great Missoula flood, or perhaps a few floods.

"It has been supposed that only a few late Wisconsin floods coursed the valleys of central and southern Washington. Bretz had argued for a single great flood (but later postulated) as many as six late-Wisconsin floods.... Judging largely from regional stratigraphy of coarse bedload deposits, Baker (1978) and Waitt (1978) acknowledged no more than three late-Wisconsin catastrophic Lake Missoula floods." [Waitt, 1980]

In 1971, R.L. Chambers and David Alt would develop a hypothesis of thirty-six floods, based on deposits in Ninemile creek. In 1980, Waitt made the case for nearly forty floods, based on evidence in the Walla Walla and Yakima valleys. Both were large departures from the commonly held view of few floods.

Chambers and Alt studied a roadcut at Ninemile creek, 20 miles west of Missoula. They found something much different than expected.

"Instead of one long sequence of light and dark glacial lake varves, we found thirty-six fairly short sequences, with layers of complexly bedded silt sandwiched between them. We interpreted the varved sequences as a record of times when the lake existed. And we became convinced that the complexly bedded layers of silt recorded a time when the lake was empty and the Clark Fork River was depositing sediment on the site.... We concluded that the lake had filled and emptied at least thirty-six times. And we counted a total of just under one thousand pairs of light and dark layers in the thirty-six sections of lake sediments. Those represented the years of Glacial Lake Missoula.

When we counted the varves between the layers of river sediment, we found that the sequence of lake sediments on the bottom of the stack contained a record of 58 years and that each successive sequence above it recorded fewer years than the one beneath. The shortest interval, nine years, was at the top of the stack....

If each successive filling of Glacial Lake Missoula involved less water than the one before, we would expect to see that pattern expressed in any flood deposits downstream. As we shall see, the oldest flood deposits in eastern Washington are the largest, and they become progressively thinner upward." [in Alt, 2002]

Chambers and Alt offered a theory based upon depositional layers from the nearby Clark Fork river and sequences deposited at the bottom of the lake. According to the theory of Chambers and Alt, the largest floods happened first, and floods successively decreased over time, presumably as the Pend Orielle lobe retreated and the height of the ice dam lessened. Likewise, the time of Glacial Lake Missoula was extremely short (geologically speaking), with the floods occurring over a span of roughly one thousand years.

Despite this startling new theory, most geologists clung to the theory of only a few floods. Not until Waitt fleshed out his hypothesis of many floods would geologists begin to debate the number of floods.

Waitt stated his hypothesis rather elegantly. "My reexamination of these rhythmic 'slackwater' deposits in the Walla Walla and lower Yakima valleys and in Pasco Basin suggests that about 40 late Wisconsin catastrophic floods deluged the region. Each rhythmite of the Tochet Beds represents a separate discrete flood that back-flooded the tributaries for hours, and was followed by decades of normal subaerial environments." [Waitt, 1980] Thus, through examination of the rhythmites deposited by each Glacial Lake Missoula flood, Wait7t came up with a much larger estimate than had been found by looking at the hydrology of Glacial Lake Missoula by Bretz and Baker..

It is worth looking at Waitt's description of rhythmites to gain some understanding into the number, and size of Glacial Lake Missoula floods.

"The measured section at the Burlingame canyon shows 39 well-formed rhythmites beneath a loess cap locally containing the Mazama tephra, the base of the section being unexposed. There are five additional thin and inconspicuous cycles near the base of the section. Thirty-nine is therefore the minimum number of recorded major events, the uncounted thin cycles representing relatively minor events. The lower 11 cycles are thicker (100 to 200 cm), have coarser bases, and some are more complex internally than are rhythmites numbered 12 through 32, which consist of regular 50 to 95 cm-thick Bouma-like sequences. The upper seven cycles are thinner and have finer sand than the central cycles." [Waitt, 1980]

This conforms with Alt and Chamber's hypothesis that the initial floods were large in magnitude, than became progressively smaller over time.

The idea of multiple layers of sediment was not new to the geological community. The explanation had always been that multiple layers formed during the same floods. Therefore, Waitt needed evidence to explain how the rhythmite layers were formed by different floods. This came from a few sources: assemblages of shells, silt on top of each layer, ash from Mount St. Helens, rodent burrows, and perhaps most convincingly, vertebrate remains. [Waitt, 1980] Waitt makes one very important observation about the number and timing of floods:

"Neither color nor texture within the Touchet Beds sections suggests subaerial exposure sufficient to produce incipient soils. There being neither conspicuous caliche horizons nor pedogenic clay at the tops of rhythmites or nonrhythmites, where the moist color is pale yellow to pale olive, the cycles must have succeeded one another quickly enough to preclude recognizable weathering." [Waitt, 1980]

This is evidence against theories which require long periods of time between successive floods. Thus, there is evidence that there must be time between floods, but it cannot be long periods of time, implying periodicity to the flood schedule.

Waitt also discusses the relationship between the evidence found in the Touchet beds and the Lake Missoula records. In fact, geologists at the University of Montana, mainly Alt, Chambers, and Curry had already suggested a hypothesis of nearly forty floods, based upon evidence found in the Missoula valley. In later work, Waitt would work to correlate the records at different locations to develop a theory which could explain sediment beds found at various locations in the flood path.

However, scientists were initially skeptical of the Waittian hypothesis of about forty floods. One issue is the idea that the ice dam had to fail catastrophically. "A variety of failures are possible,

many of which leave a considerable volume of water in the lake. At the extreme case, collapse of the tunnel roof can lead to the sweeping away of the ice dam and complete drainage of the lake." [Baker and Bunker, 1985]

Another criticism concerns Waitt's assumption that each rhythmite bed is the result of a unique flood. "(Baker) presented a 'surge' hypothesis for rhythmite emplacement in order to avoid the necessity of postulating separate floods for each rhythmite when corroborating evidence of such floods was lacking.... Indeed, such multiple beds from a single flood occur even in slack-water sediments of relatively modest-sized Holocene floods." [Baker and Bunker, 1985]

Indeed, another problem is that the number of rhythmites and varves is not the same from site to site. Although Waitt does a good job of correlating the number of rhythmites as one moves down the flood path, other have found different numbers elsewhere. "Measured sections of 62 rhythmites and 89 beds are therefore anomalous and require modification of the original hypothesis. Moreover, local unconformities in flood-slackwater sequences imply relatively long time breaks between some floods. Thus, the somewhat regular timing of floods as proposed by Waitt may be somewhat untenable." [Baker and Bunker, 1985]

In addition, massive floods are not required to form rhythmites. "It is essential to observe that flood-slackwater sediments, the rhythmites, and various lacustrine beds indicate relatively low-energy depositional environments. As recognized by Waitt (1984), the typical sequence of sedimentary structures in a graded rhythmite is accurately reproduced in a few hours of laboratory flume operation. (Ashley *et al.*, 1982). Such a sequence can be produced beneath a meter of water, as in the flume experiments, or beneath hundreds of meters of water, as proposed by Waitt (1980)." [Baker and Bunker, 1985]

One of the more important ideas that Bunker and Baker put forth is the idea that much of the flood evidence found in the channeled scablands could come from other sources, such as the Bonneville flood. This is a legitimate criticism for evidence from the Walla Walla and Yakima valleys. However, other flooding could not cause the features seen at Ninemile creek. Indeed, other floods would help to explain difficulty in correlating the number of floods based on evidence from different locations.

Indeed, Baker and Bunker propose the most important theory is their conclusions:

"The concept of tens of Lake Missoula flooding of the Columbia River should be replaced by one involving a multitude of flood events throughout the Columbia system. Many of these floods, but not all, emanated from glacial Lake Missoula. These events had variable magnitudes, some of which were truly cataclysmic. The floods followed various paths depending on their magnitudes and on the complexities of ice margins and erosional history. Although all converged on the Pasco Basin, they did so from different directions.... Considerable new work is required to accurately specify the magnitudes, frequency, and routing of the Late Pleistocene flooding in the Columbia River system." [Baker and Bunker, 1985]

Not to be outdone, Waitt answers many of the criticisms of his many flood hypothesis. Smith [1993] comes to the conclusion that Waitt's argument is convincing at the sites in the Yakima and Walla Walla valleys. Thus, the controversy appears to be settled that each rhythmite is a result of a separate flood at certain sites. However, Smith also cautions that there may be multiple rhythmites deposited in a single flood at other locations.

Waitt also makes a convincing argument for why Glacial Lake Missoula must fail catastrophically:

"The behavior of glacial Lake Missoula may be deducted from observed behavior and theoretical analyses of modern ice-dammed lakes. The glacial dams of all such lakes that lack alternative spillways are inherently metastable; many glacial lakes have drained catastrophically when the impounded water has risen deeply against the ice dam. Most or all such lakes drain before water rises enough to overtop the ice dam. There is neither field evidence nor theoretical reason that the huge

glacial Lake Missoula, whose outlet was via the ice dam, should have behaved differently from small ice-dammed lakes." [Waitt 1985]

Once the height of the water reaches sufficient height, roughly .9 the height of the ice dam, the pressure starts to pry into the seal that forms the ice dam. Once water begins to flow through tunnels created by the breaking of the seal, it generates heat, which enlarges the tunnels and eventually leads to catastrophic failure of the ice dam. [Waitt 1985] Therefore, each flood would be catastrophic, and be due to the volume of water being sufficiently large to raise lake level to a sufficient proportion of the height of the ice dam.

The height of the ice dam cannot be considered to be constant over time.

"During deglaciation, the thinning ice dam became hydraulically unstable at progressively shallower lake levels. The last many floods were therefore of much less volume than would have escaped from the maximal 2,500-km^3 lake. The flood-laid rhythmites and the Lake Missoula rhythmites are regionally thinner and finer upsection - evidence that later floods were indeed smaller (and therefore more frequent) that were earlier ones. The number of varves in the upper 15 or so Lake Missoula bottom-sediment cycles is much less than the average for all lake cycles - separate evidence that toward the end of its existence the lake dumped with increased frequency. Even half the maximum lake volume (1,250 km^3) sufficed for a mighty flood down the Columbia River valley and probably also across the Channeled Scabland." [Waitt, 1985]

Because the height of the ice dam is not constant, any modeling of the waterlines needs to include some mechanism for the changing height. This is discussed in the section on analyzing self-censored data from Glacial Lake Missoula.

The question of import for this paper is what implications this has for the famous waterline data. Considering the self-censoring aspect of the waterline data, it is reasonable to assume that most small floods were erased, as larger lake volumes would overwrite the previous records. Thus, the hills of Missoula would record the larger floods, with the records of the small floods erased as the waterlines are submerged.

3. Few Flood Theory

If Waitt's view represents a return to the principle of uniformity, then surely there must be a counterbalance. And, if uniformity is associated with the principal of late nineteenth century rationality, then perhaps a counterbalance may have a late nineteenth century religious element to it. Initially, Bretz proposed a single Glacial Lake Missoula flood, but was hesitant to propose the hypothesis due to similarity with the great flood of Genesis. Before Waitt's work, most believed that glacial Lake Missoula was caused by very few floods, possibly even a single flood. However, there is evidence to support the claim of few floods.

The single flood hypothesis starts by attacking the two assumptions on which the multiple flood hypothesis rests: "(1) Lake Missoula was the only water source for the Scabland floods; and (2) the Purcell lobe ice dam collapsed completely during each drainage event and then reestablished, resulting in lake refilling over decades of even centuries." [Shaw, 1999]

Perhaps the most difficult criticism brought forth by the single flood hypothesis is that the record left at Ninemile creek comes not from multiple floods of glacial Lake Missoula, but from other floods. "Turbidity currents generated by jökulhlaups from beneath the Rocky Mountain trench glacier to the north best explain these deposits." "We believe that deposits at the Ninemile section represent hundreds of years of normal lake sedimentation in Glacial Lake Missoula and that the thick silt beds record episodic jökulhlaups from upstream glaciers." [Shaw, 1999]

Likewise, Shaw explains the massive erosion and multiple sediment layers seen in Eastern Washington as coming from glaciers to the north, in British Columbia. Thus, Waitt's 39 layers are explained as being from multiple floods, but not from multiple floods of the Glacial Lake Missoula. "The source of the Scablands floodwater is evidently crucial to the reservoir volume and flow

duration.... (P)aleoflow directions and clast provenance in the Sanpoil arm indicate flow from the north, which is difficult to ascribe to drainage of Glacial Lake Missoula." [Shaw, 1999]

Note that the idea of a single flood is accepted in other glacial lakes. Lake Bonneville (in present day Utah) is assumed to have broken out and caused a single flood that deposited sediment among (supposed) Glacial Lake Missoula sediment layers.

The goal of this paper is not to dispute the geologic evidence for or against many floods. Rather, the goal is to determine if the waterlines on the hillsides of the Missoula valley are compatible with the hypotheses presented by geological experts. Geology has presented two alternative cases from sediment layers, one where Glacial Lake Missoula flooded many times, with the magnitude floods decreasing over time, and one with a single flood (or few floods). The goal is to use knowledge about the waterlines to determine if either hypothesis can be discounted.

4. Bathtub Rings

While most would focus only on sediment layers, there is an importance on linking the shorelines of Glacial Lake Missoula with any layers deposited:

"We can suppose that the oldest sequence of layered lake sediments corresponds to the highest shoreline. Since each successive filling lasted fewer years than the one before, the shorelines presumably become younger as they step down the mountain. All the shorelines are so faint that a bit of wave erosion would erase them. *That makes it hard to imagine that the lake level could rise above the level of an earlier shoreline without destroying it.*" (Italics mine) [Alt, 2002]

Perhaps an apt analogy could be found in a kettle of water placed on a heat source, such as a furnace, to provide humidity. A thin ring forms, due to minerals in the water, which stays after the water has evaporated. However, if the kettle should later be filled to greater volume, the original ring dissolves back into the water, and a new ring forms at the greater height.

So it is with the shorelines of Glacial Lake Missoula. Each successive filling of the lake would erase any previous waterlines of lesser height. Thus, each waterline that remains is not only lower than the others, but also younger. The highest waterlines formed first, followed by lower waterlines.

Such deposits are seen at Lake Bonneville and also Lake Agassiz in Minnesota. In Lake Agassiz, the rings are believed to be due to the lake breaking through a series of ice dams. Lake Bonneville is believed to have four main shorelines due to partial floods, with hundreds of years between floods allowing for the erosion of the shore.

In the case of Glacial Lake Missoula, the pattern is one of roughly 40 lines, with roughly equal spacing. This suggests that the height of the lake was not constant over time. It also suggests that Glacial Lake Missoula did not flood in the same pattern as Lake Bonneville, with a few major drops.

5. The Self-Censoring Data Problem

What makes analysis of waterline data particularly difficult is that it is assumed that the waterlines suffer from a self-censoring problem. This is a new problem in statistics, and requires a different approach to analyze the data.

Generally, censored data refers to data which is censored from an external source. The classic example would be data from a medical trial where patients drop out of the study before the onset of some indicator. Extensive research has been done on such data, and how to make estimates based on such data.

Rather, in this paper, self-censoring data is data where one observation affects another observation. Specifically, an observation can "censor," or delete, another observation. In the case of Glacial Lake Missoula, there is a temporal aspect to the censoring. Early observations will be censored by observations which occur later in time. One can think of self-censoring as a transformation applied to a dataset which is generated over time.

To get a handle on the concept of self censoring data, consider the following example:

5 "height" observations occur in the following order 1,2,9,3,5. An observation is censored if an observation with a higher value follows it. This would result in data as shown at each iteration.

Iteration	1	2	3	4	5
Visible Data	1	2	9	9,3	9,5

The second iteration results in a value of 2, which is greater than the value of 1, so the first observation is censored. Likewise, the third observation of 9 censors the value of 2, so the second observation (along with the first) is also censored. The fourth observation is only 3, so it does not censor the earlier value of 9. The fifth observation of 5 censors the fourth observation, but does not censor the third observation.

Specifically applying this to the Glacial Lake Missoula problem, the goal is to develop an estimate of the number of floods which occurred, based on the waterlines. However, the waterlines are assumed to suffer from the self-censoring problem. A picture is below:

The assumption for this project is that each waterline is distinct, and there is no further information to be gained about waterlines other than their height. Obviously, this is a simplification. Pardee [1910] acknowledged that the waterlines were not identical, and some waterlines are more prominent than others. However, until further research is done into the elevation of waterlines and exact thickness, the assumption will be that no further information is given from the waterlines other than the number of visible waterlines and the elevation of such waterlines.

If the waterlines (as currently observed) were a complete record of the shorelines of Glacial Lake Missoula, two very important questions about Glacial Lake Missoula should be answerable. The first question concerns the number of times the Glacial Floods occurred, since each time the lake formed a distinct waterline should form. Second, the waterlines have a perfect ordinal correlation with the volume of the lake. That is to say, the larger the volume of the lake, the greater the height of the lake and the greater the elevation of the associated waterlines.

Of course, the waterlines as currently observed do not give a complete record, due to the self-censoring aspect of the process which generates the waterlines. What the waterlines provide is an alternative method of testing hypotheses about the processes involved in Glacial Lake Missoula. The vast majority of data collected on Glacial Lake Missoula comes from the scablands of Eastern Washington, and most hypotheses are due to evidence from the channeled scablands.

However, there are advantages from correlating[1] the record of Glacial Lake Missoula to records found in the channeled scabland and elsewhere. Specifically, the waterlines, although suffering from a self-censoring problem, can absolutely be attributed to Glacial Lake Missoula. Evidence elsewhere may come from other floods. One can correlate both numbers, and magnitudes of floods to evidence found in the channeled scablands.

Perhaps the biggest advantage of the simulation is that it provides an opportunity to compare different theories as to the number and magnitude of floods to the observed waterlines. This will be further discussed in other parts of the paper.

1. The term correlation is used in the geologist's sense of matching, rather than the strict statistical sense of computing correlation coefficients.

III. Statistical Background

1. High Water Marks

Although the author believes that working with self-censored data from a likelihood perspective is difficult, some work has been done on questions of a high water marks, which relates to the idea of self-censoring data. In the tradition of "standing on the shoulders of giants," the author wishes to relate the work done on high water marks to work done on self-censoring data, to see what applies.

As defined at MathWorld, given a sequence of values $\{a_k\}$, k=1...n, the high-water marks are the values at which the running maximum increases. For example, given a sequence (3,5,7,8,8,5,7,9,2,5) with running maxima (3,5,7,8,8,8,9,9,9), the high-water marks are (3,5,7,8,9), which occur at k=1, 2, 3, 4, and 8. [Weisstein, 2004]

Beyond the similarity in name with the problems of glacial Lake Missoula, there are appear to be similarities with a self-censoring data problem. In the case of high water marks, one can consider the current observation to be censored by previous observations if previous observations are greater than the current observation. As described, the self-censored data problem is one in which current observations are censored by future observations if future observations are greater than the current observation. Thus, the problems are really identical, but run in different temporal directions.

In the case of independent, identically distributed, random variables, the expected number of high water marks / self-censored records given the number of observations can be found using combinatorics and the expected number of observations to leave a number of self-censored records can be found using combinatorics as well. If $H_n=x$ is defined where n is the number of observations and x is the number of high water marks / self-censored records, then $E(H_n)$ grows very large. For x=1,2,3,4,5,6,7,8,9,10, $E(H_n)$=1,4,11,31,83,227,616,1674,4550,12367. [Weisstein, 2004].

This has important implications for inference on self-censored data. If one assumes that observations come from the same continuous distribution, and assumes that observations are independent, then the number of observations left after self-censoring gives information about the number of original observations, *but not the distribution of those observations.* Therefore, relating the number of self-censored observations to the number of observations before self-censoring is a combinatoric problem in such cases.

Another very important property of self-censored data is that they vast majority of data will be, in fact, censored. The count[2] is an order of magnitude smaller than the number of original observations.

However, this still leaves questions to be answered. For example, even if the relationship between the number of self-censored records and number of observations (not self-censored) is known in expectations, more information about the form of the relationship such as variability is desired. In addition, the pattern of the self-censored data depends on the exact distribution. Also, when the assumption of identically distributed, independent, random variables is violated, self-censored data becomes considerably more complex to analyze.

With that in mind, the current goal of the project is to see what inferences can be made about the type of distribution, given the count and the form of the self-censored observations. The self-censoring process drastically reduces counts in a process. However, one of the interesting questions is if the form of self-censored observations can be related to the process which generated the original observations. In the case of Glacial Lake Missoula, one can observe the waterlines, which suffer from self-censoring due to the ease of erosion. However, the real interest is in the different processes that generated those waterlines. If the height of each waterline is a random variable, then the goal is to determine the distribution of those random variables.

---

[2] The term count will always be used in this paper to refer to the count of observations after a self-censoring transformation is applied.

2. Description Of R Functions

There are three main custom functions for the R programming language, a implementation of the S programming language (An Introduction To R, 2006) that were written for the analysis. The code for the functions is given in an appendix.

The first, and perhaps most important function, is the self-censoring function. This function was written to take a vector and apply self-censoring to the observations in the vector. It returns a vector of self-censored data.

The function loops over every entry in the vector. For each entry, the maximum of the entries after that entry is computed. The function then keeps the entry if it is greater than the maximum, or returns it as NA (missing) if it is less than the maximum. In the case of ties, the function keeps the latter entry and drops the former entry. The tie rule is not of importance for the vectors used in the analysis, as the entries are generated from continuous data so the probability of a tie is (almost) zero.

The vector of self-censored data is the same length as the original vector, although it may have many NA values. Additional functions such as counter.fun and extract.fun are used to grab only the desired values and remove the NA values. The counter.fun function returns the number of values which are not NA in a vector, and the extract.fun function is used to extract values which are not NA from a dataset.

Because of the looping, the function is computationally intensive. It is believed by the author that the amount of computations needed in the function increases on a polynomial power of x greater than one, where x is the length of the original vector. The computational time is negligible for vectors of length 100, noticeable for vectors of length 1,000, and beyond the author's patience for vectors of length 10,000. The function takes a few minutes when performed on a matrix of 1,000 columns of 1,000 rows each when performed on my laptop, with a 1.39 Gigahertz Celeron M process, 256 MB of RAM, running Ubuntu Dapper Drake.

A more computationally efficient method would involve looping over sorted observations, which eliminates the computationally intensive sort commands, and results in less processing. Such a function would first sort the data from maximum to minimum and consider observations to be missing if the observation number is less than that of the maximum. After the initial expense of the sort command, the function should work quickly. This is especially so when the observations come from i.i.d. (identically, independently distributed) distributions, since most will be missing. This method should be drastically more efficient, and allow research on larger datasets.

The function is also inelegant, in the sense that there is "junk" in the code which may be removable. Nonetheless, the junk is not believed to affect the computational efficiency, and it remains mainly for the author's convenience. The function also returns an error message for each vector concerning NAs, which does not prevent it from doing what is requested.

For demonstration purposes, the function is applied to the following vector (8 7 4 3 7 1 3 6 4 6). R returns the vector (8 NA NA NA 7 NA NA NA NA 6).

The second function is a random walk function used to generate the random walk data. One inputs trend and noise, and it outputs a vector of data which can be considered to come from a random walk distribution with the given trend and noise.

The general form of random walk data can be written as follows:

$$x_t=\Phi_1 x_{t-1}+\xi_t+\varepsilon_t$$

Where x_t is the observation from the current time period, ξ_t is a trend (in this case allowed to depend on time) and ε_t is an error term. If ξ is constant, then the process is called the Markov process (also known as random walk) because an observation only depends on the previous observation. [Montgomery and Johnson, 1976]

Markovian processes are a special case of autoregressive (AR) processes, processes where the current observation can be predicted from previous observations. There also exist processes where the

current observation is equal to a mean and error components from previous observations (a moving average process).

The decision was made to focus on a process where
$x_t = x_{t-1} + \varepsilon_t$ and $\varepsilon_t \sim N(0,\sigma)$,
the so called random walk with unit root. This process is interesting for a few reasons. First, as t approaches infinity, the variability of the process increases to infinity, but the expectation remains zero. Second, time series data of the unit root case (where $\Phi=1$) tend to appear in many time series when subjected to analysis. Third, adding a trend appears to be 'cheating' in the case of self-censoring. Observations with a strong trend (in comparison to variability of the error term) will either completely self-censor, or not self-censor at all. The effects of self-censoring on other AR processes (or indeed, other time series structures such as AutoRegressive Moving Average or AutoRegressive Integrated Moving Average) is an area requiring further research.

The function uses two components, trend and noise, which need to be supplied externally. The function uses cumsum to add all the previous trend components, getting the expectation for each observation. The function also uses cumsum to add all the error terms, getting the cumulative error for each observation. In practice, this creates observations with a random walk structure that can be supplied as needed. Normally, one will create noise which comes from a Gaussian distribution with mean zero and desired variability. Because the trend is supplied outside of the function, it can be chosen as needed. Note that the trend is cumulatively summed upon. Thus, if the expectation increased by one after each time period, the trend should be a vector of ones, not a vector such as (1,2,3...).

As a demonstration, the function is used to plot the four graphs given in graph III.2.1. The trend vector is a vector of 1,000 zeros and the noise vector is a vector of 1,000 observations from a standard Gaussian distribution in all four graphs.

The third custom function is used to create a uniform standardization. It works on vectors, and outputs a vector the same length as the original vector. All observations have the minimum subtracted and are then divided by the range of the data. This means that all observations in a vector will fall between 0 and 1. Also, unless all observations are the same (which will cause division by zero), there will necessarily be observations with values zero and one after the uniform standardization is applied.

The uniform standardization is advantageous for a few reasons. First, it is a linear transformation, so all properties of linear transformations apply. Second, it can be shown that the distribution of a random variable generated by a uniform standardization of another random variable is the same for all members of a scale and / or location family. Also note that the parameters of the original distribution do not have to be known to perform a uniform standardization. Third, because values are bounded between zero and one, comparisons between different distributions tend to be simplified.

This transformation is felt to better represent reality, where the scale and location parameters of the distribution of the original data are unknown, along with the form of the distribution. Using Glacial Lake Missoula as an example, transforming the elevations of the waterlines allows focus on relative distance between waterlines. Other glacial lakes would surely have different elevations, the question would be if the pattern of those waterlines is similar to that of Glacial Lake Missoula. The goal of this article is to make inference about the form of the distribution rather than the parameters of the distribution.

For demonstration, when applied to the vector (8 7 4 3 7 1 3 6 4 6), the function returns the vector (1.00 0.86 0.43 0.29 0.86 0.00 0.29 0.71 0.43 0.71), which has been rounded to two places past the decimal. The function is capable of handing missing values, and is frequently applied to self-censored vectors, which have many missing values. Note that any linear transformation of the vector would give the same vector after a uniform standardization.

With these functions, a matrix of self-censored random variables can be constructed. One initially creates a matrix with random variables from whichever original distribution is desired, such as

uniform, exponential, normal, or random walk with each column a different simulated sample. One than applies the self-censoring function to get simulated self-censored data from the appropriate distribution. Likewise, one can apply the uniform standardization, for reasons discussed in the description of the function. The end result is a matrix with observations from the self-censored transformation of the original distribution.

3. A Short Bit About Statistical Properties Of Self-Censoring Datasets

In the case of Glacial Lake Missoula, there is little controversy over any scale-location parameters describing the distribution of the heights of the lake. Thus, the main thrust of this paper is not to make inference about scale-location parameters after self-censoring, but rather counts and general forms of distributions (specifically time dependent versus non-time dependent).

For statistical purposes, it is worthwhile to say a bit about statistical properties of self-censored data.

In any self-censored data set, there are two observations which are not affected by self censoring. The first is the maximum, which will not be censored by future observations. Likewise, the minimum of the data set will necessarily be the last observation. However, this is not an unbiased estimate of the minimum, but rather an observation from the original distribution which is unaffected by the self-censoring transformation.

Thus, if one believes that data has undergone a self-censoring transformation, one has two statistics to use which are unaffected by the self-censoring transformation, the maximum and the last observation.

4. Statistical Analysis Goals

The main goal of this paper to consider ways of generating data; and what data will look like after a self-censoring mechanism is applied. For example, how does data with an explicit temporal structure (either a trend or a random walk) look after self-censoring, as compared to observations which do not have a temporal aspect?

When examining a self-censoring transformation, a few questions come to mind:

1. What does self-censored data from each of these distributions look like. That is to say, do certain patterns (such as banding) emerge from self censored data?
2. What is the relationship between the count and number of uncensored observations?
3. How much variability is there in the count, given the number of uncensored observations?
4. Are those relationships dependent on scale / location parameters for distributions of a scale location family, after the data have been suitably standardized?
5. Can one determine which distribution self-censored data comes from, given self-censored data? That is to say, does the pattern of self-censored data that comes from a uniform distribution look different than the pattern of self-censored data that comes from a normal distribution?

While no geologist which I am aware of has suggested that the height of Glacial Lake Missoula comes from some sort of i.i.d. distribution, these questions are important for statistical purposes.

For observations which have some sort of temporal aspect (either trend or random walk), the same questions are also of interest. Also, of particular interest is how one estimates parameters from self-censored data. For the time being, the focus will be on discerning the form of the distribution more than estimating parameters of a distribution from self-censored data.

Note that research already performed on high water marks answers some of these questions. The number of observations after self censoring depends only on the number of observations before self censoring, for any continuously distributed random variable. The form of the distribution does not matter. Thus, the number of self-censored observations does not give any information about which

independent and identical distribution those self-censored observations came from. Thus, the second through fourth questions are actually questions of combinatorics for variables which are i.i.d., which have already been answered by research on high water marks.

However, one of the goals of this project is to determine if self-censored data come from an i.i.d. distribution or a distribution with a temporal structure, and the form of that temporal structure. The number of observations will be useful for solving such questions.

Another goal is to see if one could differentiate between i.i.d distributions after a self-censoring transformation is applied. The number of observations gives no additional information. Also, a transformation such as a uniform standardization should be used to see if the form of distributions is different rather than just scale and location parameters.

One option would be to use a suitable test along with simulated self-censored data. Consider a two-sample test such as the two-sample Komogorov-Smirnov test. Under a hypothesis test, one would expect to reject the null hypothesis with probability α if the simulated self-censored data come from the same underlying distribution. By examining the simulated probability of rejecting a null hypothesis for self-censored data, one can get an idea how different the self-censored distributions are. If the distributions are different, after self-censoring, then one would expect such a test to have high power.

The Kolmogorov-Smirnov test for two samples is appropriate in this case. The hypotheses of interest are:

H0: $F(x)=G(x)$ for $-\infty < x < \infty$

H1: The hypothesis H0 is not true [DeGroot and Schervish, 2002]

The advantage of the Kolmogorov-Smirnov test is that it simply asks if the two c.d.f.s (cumulative distribution functions) are the same for two populations. The K-S test looks at the largest difference in the empirical c.d.f.s to determine if the two samples come from populations with the same c.d.f.s.

Another important question concerns differentiation between self-censored data that originally came from an i.i.d. process versus data which came from a random walk or trended process. That is, is there some sort of temporal structure to the process? As mentioned before, the number of observations is useful in such analyses. Thus, if one considers the number of observations from a self-censored dataset as coming from some distribution, the question is if those distributions are different for different types of data. Further, how different are the distributions of the counts of self-censored data. Thus, the goal in this case is to look more at the counts, since the count gives information related to this aspect of the process.

5. An Initial Examination Of Empirical Cumulative Distribution Functions

In any analysis, the first place to start is with a graphical analysis. With the matrices given, it is relatively easy to use graphical tools to analyze the forms of self-censored data.

One tool is a dotplot, which simply plots each observation in a dataset as a dot along an x-axis. A dotplot is relatively useful for gaining information about the number of observations in a distribution, and one can get a relative idea of where those observations occur.

One thing that is observed is random walk data tends to have far more self-censored observations than data that comes from any i.i.d. distribution. Thirteen out of the sixteen graphs show enough observations to be difficult to count, whereas they are countable in any of the i.i.d. distributions. However, one graph shows a mere two observations.

Perhaps the most useful tool is examination of the empirical c.d.f.s of the different distributions. An empirical c.d.f. can be defined as follows: "for each number $x(-\infty < x < \infty)$, the value of $F_n(x)$ is defined as the proportion of observed values in the sample that are less than or equal to x." [DeGroot and Schervish, 2002] One property of the empirical c.d.f. is that $F_n(x)$ converges in probability to $F(x)$ for $-\infty < x < \infty$ as n approaches ∞. It is not known if this is true for samples in which a self-censoring transformation is applied. Not only is there the problem of the count not necessarily increasing with n,

but the transformation itself may not give a convergent empirical c.d.f., even as the count approaches ∞.

Whereas the dotplot gives more information about the number of self-censored observations, the empirical cdf gives more information about the form of the self-censored distributions. These are given in plots III.5.1, III.5.2, III.5.3, and III.5.4.

It does appear that the c.d.f. for self-censored observations from an exponential distribution appears to have more jumps in the left, implying relatively smaller values tend to be seen from exponential data. This is likely due to the strong right skew of the exponential distribution, which means that the maximum is likely to be much larger than all other values in the sample. This is generally the same pattern that an empirical c.d.f. would show without self-censoring, with most of the jumps in the right part. This is similar to the theoretical c.d.f., which looks like an inverted exponential p.d.f.

The self-censored normal data appears to show less of a pattern in jumps. The self-censored uniform data tends to show jumps in the right, likely due to lack of outliers.

This contrasts with the empirical c.d.f. for the self-censored random walk data. They generally show a trend of forty-five degrees in a roughly straight line. This is the c.d.f. of a uniform random variable (without self-censoring). The author has no explanation why the c.d.f. should appear in such a form.

One area of future research which may prove fruitful would be to model the different empirical c.d.f.s as c.d.f.s of beta distributions. Such modeling may give more information about the structure of self-censored data.

With this in mind, the next goal is to develop a quantitative analysis of self-censored data.

6. Predicting The Distribution From The Count

Based on the graphical tools and considerations of data analysis, there appear to be two paths to take. First, differentiation between random walk and i.i.d. data appears to be possible based on counts, as the counts from random walk data appear to be much higher. Second, differentiation between self-censored data from different distributions does not depend on the count, but does depend on the form of the empirical c.d.f. of the self-censored data. With this in mind, techniques for both types can be developed.

Considering that the goal is to develop techniques to identify the form of the data, given the count, logistic regression seems to be a logical choice. Logistic regression is used when the response variable falls into two categories, success and failure. It models the probability of success given a set of predictors. The logistic regression equation can be written as

$P(X=1)=1/(1+\exp(\beta_0+\beta_1 x))$ [Munro, 2005]

Using logistic regression is particularly appropriate, because in reality, one will have a count of observations after censoring, but will not know if there is a temporal dependence in the original data structure. Therefore, being able to use the count as a tool is important. As before, the particular focus is on data with a pure random walk structure, and no trend, but results could easily be found assuming other structures.

Logistic regression is used to predict whether the self-censored data comes from an i.i.d. or random walk distribution. As explained elsewhere, the choice of i.i.d. distribution does not impact the count. 1,000 observations of length 1,000 are simulated from both exponential and random walk distributions. The self-censoring function is applied to each observation. Then, the extract and counter functions are used to get 1,000 counts from exponential distributions and 1,000 counts from random walk distributions, after self-censoring is applied. These counts are then used to predict if an observation came from a random walk distribution or an i.i.d. distribution through logistic regression.

The i.i.d. distribution used for these results was the exponential distribution, but the distribution does not matter.

As a trial, it was decided to simply use 1,000 observations in each data set before self-censoring was applied. This was done for two reasons. The first is that it is a round number, and seems a reasonable place to start for differentiability between different processes. The second is that Alt and Chambers hypothesized that Glacial Lake Missoula lasted for roughly 1,000 years, based on a count of 1,000 varves at Ninemile creek. [Alt, 2002] Thus, the use of 1,000 observations has a Glacial Lake Missoula connection, with each observation representing the height of the lake in a calendar year.

The logistic regression allows one to make predictions of the type of process involved in the original dataset, based on the counts. A boxplot III.6.1 is constructed of 1,000 datasets of 1,000 observations each, with 1,000 being generated from exponentially distributed data and 1,000 being generated from random walk data. Looking at the boxplot, one can see that the largest count from the exponential data is roughly 20, which is roughly the first quantile of the counts from random walk data. The counts from the random walk data generally tend to be much larger. A five number summary for the exponential data is (1,6,7,9,17) and a five number summary for the random walk data is (1,13,27,47,140). Because of the lack of overlap it appears that one can well predict the form of the original data based on the count after a self-censoring transformation is applied.

After examination of the boxplots, the next step is to use logistic regression to make predictions. The simulated data is used to fit a logistic regression curve. 1,000 observations should be sufficient to fit a curve that is reasonably close to the population curve, although one could generate more samples if there is concern about variability in estimates of the curve. This would be useful in cases where the prediction probabilities are close to .5, and more precision in the curve results in more precision in the estimate.

Predictions for each count are plotted in plot III.6.2, along with rugs of the counts from self-censored data from the exponential distribution on the bottom, and counts from self-censored data from a random walk on the top. Each count is predicted to be from either the exponential or random walk distribution, based on the probability given from the logistic regression, with a prediction cutoff point of .5. Of the 1,000 counts from exponential data, 965 were correctly predicted to be from an exponential distribution. Of the 1,000 counts from the random walk data, only 764 were predicted to be from a random walk distribution. This is likely due to most low counts simply being assigned to the exponential distribution, which makes sense given the very large number of low exponential counts. However, using logistic regression to predict which distribution the data comes from based on the count does give more power than random guessing.

The analysis was also done using only 100 original observations. A five number summary for the self-censored counts from the random walk data is (1,5,10,16,41). For the exponential data, the five number summary is (1,4,5,6,12). However, the overlap between the two distributions is much larger, and the logistic prediction was accurate for 884 and 612 of the 1,000 samples from exponential and random walk distributions, respectively. Histogram III.6.3 is made of the counts for the random walk data, and also for the exponential data. In both cases, the counts take a right skewed distribution, but the strength of the skew is stronger in the case of the random walk data.

7. Predicting The Count For Random Walk Data

Although the relationship between the original number of observations and the count after self-censoring is known for i.i.d. variables, less is known about the relationship for temporally dependent data.

As before, a matrix is constructed of random walk data. From this, columns of differing lengths are extracted and the self-censoring function is applied, then the count examined. Thus, one gets a vector of differing numbers of original observations and the count after self-censoring is applied. For simplicity, the number of original observations ranges from 1 to 100, with five repetitions at each number of original observations. Thus, there exist 500 pairs of points of data, with values of the

original number of observations varying from 1 to 100, and a count after the self-censoring transformation is applied. The original number of observations and the counts are plotted in scatterplot III.7.1 Kernel smoother and linear regression lines, with the count as a function of the number of original observations are added.

The estimated slope coefficient of the linear regression line is roughly .10, and the estimated intercept is 2.65. The R-Squared of the model is only .2386, so there is much variability that cannot be explained. It does appear that the linear form does not hold, so perhaps the count is related logarithmically or with a square root transformation to the number of observations. If a logarithmic transformation of the count is taken, the intercept estimate is .948, the estimated slope coefficient is 0.148 and the R-Squared is .2235.

To verify this, the program was rerun, this time with the number of original observations varying from 1 to 1,000. While far more computationally expensive, due to having more observations to apply the self-censoring function and also due to the computational expense of applying the self-censoring function as the number of observations increases, it is useful to see if the estimated slope coefficients are stable over a wider range of original number of observations.

Due to concern over computational expense, only two repetitions were used. When the number of original observations ranges from 1 to 1,000, the estimated linear regression line has intercept estimate 9.09 and slope estimate of .028, with an R-Squared of .1842. If a log transformation is applied to the response variable, the estimated linear regression line has intercept of 2.018, slope of .001437 and R-Squared of .1609. This is seen in scatterplot III.7.2.

Interestingly enough, neither the slope coefficients from the linear or logarithmic model are stable as the number of observations increases. This is odd, as one might expect that the count would be well modeled as an exponentially increasing function of the original number of observations. However, the exact form appears to be more complex, even when transformations are applied. Likewise, the relationship appears to weaken as the number of original observations increases. Thus, while the expected count increases as the number of original observations increases, there may be little more than can be meaningfully gleaned.

Due to computational expense, simulating for more than 1,000 original observations is not performed. However, the regression lines and plots give one an idea of how the count after self-censoring varies with the number of original observations for random walk data.

8. The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) is a non-parametric test of whether two samples come from populations with the same cumulative distribution function, through examination of empirical cumulative distribution functions, which come from samples. Because of the general nature of the alternative hypothesis, the test is applicable to a question of rather there is a difference in the form of the data after self-censoring is applied.

The empirical c.d.f.s are defined in section III.5 as $F_{n1}(x)$ and $F_{n2}(x)$. The two-sample, two-tailed test statistic is defined as:

$D=\text{maximum} \mid F_{n1}(x) - F_{n2}(x) \mid$ [Daniel, 1990]

For small samples, it is usually best to find the exact p-value. There are (n_1+n_2) C (n_1) (where the C is used to signify choose) possible ways to assign the two groupings to the sample data. Each assignment gives a different value of the test statistics. The proportion of those possible values which are exceeded by the test statistic gives the p-value.

The K-S test statistic is well defined, even in the presence of ties. However, R returns an error message if two vectors which contain ties are used. According to the R help files, "the presence of ties generates a warning, since continuous distributions do not generate them." [R Help, 2003] However, in this study, a transformation is used which guarantees two ties between distributions, as both vectors will have the values of 0 and 1.

The solution that was used was to jitter (add a small amount of noise) one of the samples in R to prevent ties. According to the R help files, "'jitter(x,...)' returns a numeric of the same length as 'x', but with an 'amount' of noise added in order to break ties." [R Help, 2003] It is common to jitter data when ties would be problematic.

9. Use Of The K-S Test On Uncensored Data

One goal of any test is to relate the power of the test to the significance of the test. Recall from introductory statistics that a test has probability α of rejecting the null hypothesis, when it is in fact true (significance, the probability of a type I error); the same test has a probability of $1-\beta$ of correctly rejecting the null hypothesis if an alternative hypothesis is true (power, 1- the probability of a type II error). As one increases the level of significance, power increases.

Ideally, a test will have higher levels of power than the level of significance for a given alternative. Likewise, if the null hypothesis is in fact true, one would expect the power of a test to be exactly equal to the level of significance chosen. The goal of this project is to compare the power to the level of significance in the two-sample K-S test as distributions under transformations are selected. For example, if one were to perform a K-S test where both samples were generated from a uniform distribution, because the null hypothesis is true, the probability of rejection (of the null hypothesis) of such a test should be equal to the level of significance chosen. However, the probability of rejection may not be the same as the level of significance if a uniform standardization or self-censoring transformation is applied, even if the data is generated from the same original distribution.

First, 8 non-censored observations were taken from the uniform and exponential distributions. 8 is roughly the expected count after censoring for 1,000 original observations. The uniform standardization was applied to both sets of observations. The empirical c.d.f. after the transformation has been applied is plotted for both sets of observations. The K-S test finds the largest difference in the two empirical c.d.f.s to be D=.35. From this test statistic, the p-value (based on all possible differences) is calculated to be .7692. The two e.c.d.f.s are seen in plot III.9.1.

Of course, because the data is simulated and can be gathered extremely cheaply, one approach is to compute power for different levels of significance, by simulating a large number of samples. 1,000 K-S tests are performed on simulated data, and the p-values from the tests are extracted. For levels of significance ranging from 0 to 1, the proportion of the 1,000 tests which have p-values less than levels of significance are calculated. The proportion for each level of significance is plotted on the y-axis, with the levels of significance on the x-axis, and a 45 degree line is added. This is similar to a receiver operating curve (R.O.C.), where the goal is to compare false positive and false negative rates.

If one examines the exponential versus the uniform distribution, in plot III.9.2, the power generally remains below the chosen level of significance, although it does increase, implying that there is power to detect differences even when the uniform standardization is applied. The reason why the power is generally below the level of significance is because the p-values are often conservative for small sample sizes.

One option is to compare power and significance for the same distribution. Because the null hypothesis is known to be true, the probability of rejection should be the same as the level of significance, for all levels of significance. When an exponential distribution is compared to another exponential distribution, the probability of rejection retains the stepwise aspect. This is in the plot III.9.3. The probability of rejection is below the chosen level of significance, for any level of significance.

The distributions were also tested against random walk data, which was given 27 observations based on results comparing counts. In this case, power is a concave function of significance, implying that the K-S test is useful for comparing the distributions. This is after the uniform standardization

function was applied to the random walk data. So, the shapes do differ enough for detection if the sample size is increased. This is seen in plot III.9.4

However, when the sample size was decreased to 8 for the number of random walk observations the power is usually below the level of significance, as when comparing exponential and uniform distributions. This is seen in plot III.9.5.

Based on the above plots, the K-S test does appear to have some power to differentiate between distributions, even after a uniform standardization is applied.

10. Use Of The K-S Test On Self-censored Data

Because the K-S test appears to be useful in comparing distributions, even with small sample sizes, the next goal is to use the K-S test to compare data, which has undergone a self-censoring transformation. Thus, any data will first undergo self-censoring, and then a uniform standardization will also be applied.

As before, 1,000 original observations were repeatedly generated from different distributions (exponential, uniform, normal, random walk), then the self-censoring transformation was applied, leaving counts that were generally in the single digits, except for the random walk data. For computational efficiency, the decision was made to generate only 200 sets of original observations for each distribution, then randomly sample from each set (with replacement) to perform the K-S test. Thus, a K-S test might be performed on the 20th set of observations from the uniform distribution and the 139th set of observations from a random walk distribution. As before, 1,000 such tests were performed.

When comparing data from an exponential to exponential distribution, the probability of rejection of the test was always found to be less than the level of significance. One potential reason for the low probability of rejection may be due to extremely small sample sizes. Another reason could be the decision to jitter the data, and the artificial ties induced by the uniform standardization transformation. Because the same transformations are applied to data which is generated from the same distributions, the population c.d.f.s should be the same after transformations. Similar results were found when comparing data from normal distribution to normal distribution. Plot III.10.1 has the proportion of tests which are rejected on the y-axis, and the chosen level of significance (α level) on the x-axis is given, using the normal distribution.

However, the results found when comparing uniform to uniform, (plot not given) In this case, the probability of rejection was found to be nearly zero, regardless of the level of significance chosen. In other words, the test is returning p-values near one nearly all the time.

Similar results were seen when comparing random walk to random walk distributions, which are given in plot III.10.2. The low probability of rejection when comparing random walk distributions is odd, since the counts are generally much higher when the original distribution is random walk. Thus, there is something in the structure of the data which makes the K-S test give large p-values. Exactly why the K-S test does not work after applying a self-censoring transformation to random walk data is an area for further research.

When comparing different distributions, the same problem arises. Regardless of which distribution is compared to either random walk or uniform data, one gets a graph similar to that shown in the previous paragraph. If one compares normal to exponential data, then one gets a graph similar to that when exponential is compared to exponential.

One criticism of the use of the K-S test on self-censored data is that the lack of power may be an issue of very small sample sizes or the uniform standardization. Perhaps the best evidence that the self-censoring transformation is causing the low power is found by examining graph III.10.3 which compares power to significance after self-censoring transformations have been applied. Recall that power was higher than significance when comparing 8 observations generated from the exponential distribution and 40 observations generated from a random walk distribution and a uniform standardization was applied. Counts of 8 and 40 would not be unusual to see after a self-censoring

transformation is applied to 1000 observations generated from the respective distributions. When simulating data from the respective distributions, and applying a uniform standardization, the K-S test is seen to be effective. However, when data is simulated from the respective distributions and self-censoring is applied to get counts roughly equal to those found when simulating from the known original distributions, then a uniform standardization applied, the K-S test nearly universally returns p-values near 1. This is important, because it eliminates small sample size and uniform standardization as reasons why the K-S test returns high p-values. Thus, the self-censoring transformation alters the data structure in such a manner that the K-S test becomes unable to differentiate between distributions.

Perhaps one reason why the K-S test fails to work when a self-censoring transformation is applied is due to the influence of the last observation on the uniform standardization. All observations, after self-censoring, are bounded between the maximum and the final observation. Because of this, there is much variability in the lower bound of the standardization.

With further research, perhaps methods which can distinguish based on the form of the e.c.d.f. will be developed. Nonetheless, as shown in the first section, the count is an extremely useful tool for determining if data was generated from a distribution with a temporal structure.

IV. Geological Conclusions

1. Analysis Of Glacial Lake Missoula Theories

Even with a description of self-censored data, the problem seems to be a rather difficult one to answer. The best description for self-censored data is "time dependent order statistics." [Banfield, personal comm.] Thus, each visible observation is known to have occurred after visible observations which are larger than it.

In the case of Glacial Lake Missoula, one can view the height of different waterlines as coming from different processes based on different theories. Based on different theories, one can simulate data. One can then apply self-censoring to the simulated data and compare the simulated data (after self-censoring) to the observed data.

Going back to the classic equation of model building, DATA=FIT+RESIDUAL. Each theory suggests a different number of floods (or original number of observations), along with a different model for the FIT portion of the data. Likewise, it seems necessary to include a residual portion to incorporate the imperfections of each theory in adequately describing complex processes.

Consider the theory advocated by Waitt and others that Glacial Lake Missoula was formed from a lobe that came down from the Cordilleran ice sheet, and that Glacial Lake Missoula flooding was generally catastrophic in nature. Under this theory, the volume of Glacial Lake Missoula decreased as the Cordilleran retreated.

The height is modeled as a random walk, with a strong trend component. Even though the dam failed and reformed, along with the lake, it is reasonable to assume that the height at each forming is related to the height at the previous forming.

Note that under this scenario, the self-censoring aspect of the shorelines is of little concern, as self-censoring will have no effect if observations are naturally decreasing. Whereas all information is lost if observations are naturally increasing, all information is kept if observations are naturally decreasing, except in those cases where a later observation is greater than one or more observations, possibly due to either a change in the process, or the natural variability present in the process.

The situation is different if one believes that multiple waterlines can be left by a single formation of the lake. In such a case, it seems reasonable to believe that such waterlines would come from a random walk distribution. Although Glacial Lake Missoula had no outlets (other than the ice dam), effects from intake and evaporation mean that the height was unlikely to be constant. It seems reasonable to model the height as a random walk, as the expected height would be equal to the height of the lake from last year, plus noise from the combined effects of intake and evaporation.

Thus, three main hypotheses for flooding are brought forth. The first says that the height of each watermark comes from a distribution with a very strong trend component, relative to the error from year to year, and that there were forty floods. This is consistent with Waitt's hypothesis. In this case, the trend will be considered to be twice the error term. However, the data will still be considered to come from a random walk distribution rather than one of pure trend. This is because the height of the ice dam should be thought to be related from flood to flood. The second differs in that there were 89 floods. This is consistent with Atwater's estimates. The third hypothesis is that there was only one flood, but the height of the lake varied from year to year according to a random walk process. Thus, one can view the waterlines as self-censored data from 1,000 random walk observations.

When constructed as above, the number of waterlines from the Wattian theory varies between 35 and 40 when 1,000 simulations are done. The number of waterlines as suggested by Atwater's count varies between 82 and 89 when 1,000 simulations are done. Note that the number would decrease if the variability of the error term were increased relative to the strength of the trend. Thus, that is not to say that one would fail to see waterlines as seen on the hillsides of Missoula if 89 floods did occur, but

rather one would have to justify the strength of the error term relative to the trend. This is seen in histograms IV.1.1 and IV.1.2.

Unexpectedly, the range of counts from a random walk process does include the roughly 40 waterlines seen on the hillsides of Missoula. The exact distribution is discussed in much more depth in section III.6. Considering that the median number of "waterlines" observed in the simulation was 27 and the third quartile was 47, a value of 40 waterlines cannot be considered unusual.

Thus, the count will not determine if the heights of the lake come from a distribution with strong trend and few observations or else weak trend and many observations. One option would be to examine the empirical c.d.f. of such a distribution, which has been shown to be useful when the sample sizes are reasonably large. However, based on work with the Kolmogorov-Smirnov test in simulation, such as test may not be able to distinguish differences when self-censoring transformations have been applied. An alternative, which is beyond the scope of the current investigation, is to use other techniques which relate the differences in height to different processes (differencing), and examining the differences using time-series analysis techniques.

However, hopefully examination of the waterlines from Glacial Lake Missoula will provide insight into any other theories about the processes of Glacial Lake Missoula. It seems odd to come away with a conclusion that there is not enough evidence to resolve which view is more accurate based on a new technique. Nonetheless, refinement in existing theory will allow one to develop more accurate estimates on which to apply self-censoring.

V. Conclusions And Suggestions For Further Work

1. Suggestions For Further Work

As with any project, there remains much to be done. The following are useful to study the impact of a self-censoring transformation.

One question of note would be estimation of scale and location parameters of the original distribution when given data that has undergone a self-censoring transformation. In particular, are observations other than the maximum and last observation useful for estimation?

Limits and self-censoring would also be another question of interest. It can be proved that the empirical c.d.f. does converge to the population c.d.f. in a variety of situations, particularly the case of i.i.d. random variables. Less is known about convergence if a self-censoring transformation is applied. Perhaps the empirical c.d.f. does not converge to anything after a self-censoring transformation is applied.

Another option with the K-S test would be to test original distributions against the distribution under self-censoring. The graphs of the empirical c.d.f.s appeared to take an exponential shape. Therefore, it may be possible that certain distributions retain shape under a self-censoring transformation.

More insight as to the relationship between the count and number of original observations is an area of further research. Most of the research focuses on random walk data, but there are a variety of ARIMA processes which could be investigated. In a random walk process, each observation is equally likely to fall above or below the one that proceeds it. If that probability is altered, how does the self-censoring transformation affect the questions asked earlier in the paper?

For observations generated from a random walk process, the relationship between the count and the number of original observations is unknown. It appears to be neither linear, nor exponential, based on lack of stability when moving from 100 to 1,000 observations. The appropriate functional form is still unknown. More work is needed to find which form to use in the relationship. Mark Greenwood [personal comm.] has suggested the use of Box-Cox methods to suggest a suitable transformation of the count to where the relationship with the number of original observations is linear (under transformation).

Further research into high water marks would dovetail with the results of this paper. The main use of research into high water marks was to justify that counts would be the same from any i.i.d. distribution. However, with more understanding of the combinatorial arguments, more work can be done from a likelihood perspective. For example, the explicit probabilities of getting counts, given the number of original observations, can be computed.

Another goal would be to use a specific dataset of lake heights, and attempt to correlate that dataset to the results from the simulation. One could use the true dataset to perform the Kolmogorov-Smirnov testing, in an attempt to determine compatibility with data simulated from different processes.

Also, perhaps this will spur the geological community to do further research into the waterlines that remain on the hillsides of Missoula. Most of the work has moved further and further downstream. Shaw brings a very important criticism, which is that downstream evidence may not be tied to Glacial Lake Missoula. However, no one questions that the waterlines were left by Glacial Lake Missoula. Further research into the thickness of the waterlines and attempts at dating such waterlines would give volumes of information about the lake. Likewise, any theory about the Glacial Lake Missoula floods should at least give some consideration to the lines on the hill. Correlation of evidence from different sources will give more accurate estimates as to the number, and the magnitudes, of the floods.

2. Conclusions

To come away after many hours of research and conclude that both a single flood and multiple flood hypothesis are compatible with the waterlines of Glacial Lake Missoula initially seems quite unsatisfying. However, it seems quite incredible that such different processes can result in the same results. In one process, the lake only has to flood forty times to leave forty waterlines. In another process, the lake floods once, leaving thousands of waterlines, and yet only forty remain after the self-censoring transformation. That radically different datasets give rise to the same data after the transformation is applied really is a discovery. I was initially skeptical that a one-flood theory could explain the nearly forty waterlines seen on the hillsides of Missoula.

Likewise, the sheer number of observations required to get relatively small counts after self-censoring was quite surprising. To start with a dataset of one thousand observations and have a transformation that reduces it to an average of eight observations is unexpected, to say the least. The required number of observations appears to grow at an exponential rate, which is something that was not suspected before the project.

Appendix 1: R Functions

self.censor.R.txt

```
####################################33
#Creates Self-censoring function
#Gives self-censored dataset with
#NA for missing values.
#Works as a source.
####################################
####################################
#BEGIN FUNCTION
####################################3
self.censor <-function(obs)
{
n.iter=length(obs)
time <-1:length(obs)
########################################
#RUN THIS STUFF FIRST!
########################################
block=NULL
cen.obs=1:n.iter
future.obs =1:n.iter
future.big=1:n.iter
########################################


#Run this once for intialization!
#Input vector "obs"
#Returns self-censored vector "cen.obs"
#Uses numeric value 999 for censored observations.
################################
for (iter in c(1:n.iter))
{#Open 1
##############################
#Works when given iteration
future.nums <-(iter+1):n.iter
future.big <-max(obs[future.nums])
cen.obs[iter] <- ifelse(obs[iter] > future.big ,
obs[iter],"N/A" )
#999 for missing data.
#Code to restore last observation
}#Close 1
cen.obs[n.iter]=obs[n.iter]
cen.obs <-as.numeric(cen.obs)
##############################
}
```

counter.fun.R.txt

```
#Cody L. Custis
#Function Written By Jim Robison-Cox
counter.fun <-function(x){sum(!is.na(x))}
# A function to count values which are not missing in a vector
```

etract.fun.R.txt

```
#Cody L. Custis
#Function Written By Jim Robison-Cox
extract.fun <-function(x)
{
x[!is.na(x)]
}
```

```
# A function to extract values which are not missing in a vector
```

random.walk.R.txt

```
#Cody L. Custis
#R Code To Generate A Random Walk

random.walk <-function(trend,noise){
cumsum(trend)+cumsum(noise)}
```

stand.unif.R.txt

```
stand.unif <-function(x){ #Open 1
vec.range <-max(x,na.rm=T)-min(x,na.rm=T)
(x-min(x,na.rm=T))/vec.range
}#Close 1
```

Appendix 2: Selected R Programs

Comparison.R.Win.txt

```
#Cody L. Custis
#Comparison.R.Win.txt
#Used to compare counts in walk and exponential distributions.
source("E:/Glacial Lake Missoula Project/GLM Project Summer/random.walk.R.txt")
source("E:/Glacial Lake Missoula Project/GLM Project Summer/stand.unif.R.txt")
source("E:/Glacial Lake Missoula Project/GLM Project Summer/counter.fun.R.txt")
source("E:/Glacial Lake Missoula Project/GLM Project Summer/extract.fun.R.txt")
source("E:/Glacial Lake Missoula Project/GLM Project Summer/self.censor.R.txt")
#Loads stand.unif,self.censor,random.walk
set.seed(181845)
n.samples <-1,000
l.samples <-1,000

walk.var=matrix(c(1),nrow=l.samples,ncol=n.samples)
for (iter in 1:n.samples)
{
trend <-rep(0,l.samples)
noise <-rnorm(l.samples)
walk.var[,iter] <-random.walk(trend,noise)
}

exp.var=matrix(c(1),nrow=l.samples,ncol=n.samples)
for (iter in 1:n.samples)
{
exp.var[,iter] <-rexp(l.samples)
}

#Be patient, may take a long time.
censored.exp <-apply(exp.var,2,self.censor)
censored.walk <-apply(walk.var,2,self.censor)

num.exp <-apply(censored.exp,2,counter.fun)#
num.walk <-apply(censored.walk,2,counter.fun)

bart.exp <-cbind(as.numeric(num.exp),0)
summary(bart.exp)
#
#       X1               X2
# Min.   : 1.00   Min.   :0
# 1st Qu.: 6.00   1st Qu.:0
# Median : 7.00   Median :0
# Mean   : 7.45   Mean   :0
# 3rd Qu.: 9.00   3rd Qu.:0
# Max.   :17.00   Max.   :0
bart.walk <-cbind(as.numeric(num.walk),1)
summary(bart.walk)
#       X1                X2
# Min.   :  1.00   Min.   :1
# 1st Qu.: 13.00   1st Qu.:1
# Median : 27.00   Median :1
# Mean   : 32.95   Mean   :1
# 3rd Qu.: 47.00   3rd Qu.:1
# Max.   :140.00   Max.   :1
bart <-rbind(bart.exp,bart.walk)
bart <-as.data.frame(bart)
bart$V1<-as.numeric(bart$V1)#Gives what is wanted

#Creates boxplots for comparison
par(mfrow=c(1,1))
pdf("E:/Glacial Lake Missoula Project/GLM Project Summer/comp.boxplots.pdf")
boxplot(V1~V2,data=bart,xlab=c("Exp=0","Walk=1"))
dev.off()
```

```
logit.bart <-glm(V2~V1,family=binomial(link="logit"),data=bart)
preds.bart <-predict(logit.bart)
probs.bart <-1-1/(1+exp(preds.bart))
#Predictions Graph
pdf("E:/Glacial Lake Missoula Project/GLM Project Summer/predict.pdf")
plot(bart$V1,probs.bart)
rug(bart$V1[1001:2000],side=3)
rug(bart$V1[1:1,000],side=1)
dev.off()


#Test Accuracy
right.exp <-probs.bart[1:1,000]
good.exp <-right.exp[right.exp<.5]
length(good.exp)#965

right.walk <-probs.bart[1001:2000]
good.walk <-right.walk[right.walk>.5]
length(good.walk)#764

#Summary Of Logit Model
summary(logit.bart)

#Call:
#glm(formula = V2 ~ V1, family = binomial(link = "logit"), data = bart)
#
#Deviance Residuals:
#    Min       1Q   Median       3Q      Max
#-1.6286  -0.6843  -0.1766   0.2603   2.3686
#Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
#(Intercept) -2.97777    0.14387  -20.70   <2e-16 ***
#V1           0.23502    0.01335   17.61   <2e-16 ***
#(Dispersion parameter for binomial family taken to be 1)
#
#    Null deviance: 2772.6  on 1999  degrees of freedom
#Residual deviance: 1524.1  on 1998  degrees of freedom
#AIC: 1528.1
#Number of Fisher Scoring iterations: 7
```

Count.Orig.R.txt

```
#Cody L. Custis
#Count.Orig.R.txt
#Used to create graphs of number of original observations and count after self
#censoring
source("/home/gecko/MSUDocs/GLMProj/GLMProjSum/useful.R.txt")
#Loads stand.unif,self.censor,random.walk
#Set up to use five reps

n.samples <-100
l.samples <-100

walk.var=matrix(c(1),nrow=l.samples,ncol=n.samples*5)
for (rep.lisa in 0:4){
set.seed(100*rep.lisa+365)
for (iter in 1:n.samples)
{
trend <-rep(0,l.samples)
noise <-rnorm(l.samples)
walk.var[,(iter+(n.samples*rep.lisa))] <-random.walk(trend,noise)
}
}

cen.var=matrix(NA,nrow=l.samples,ncol=n.samples*5)
```

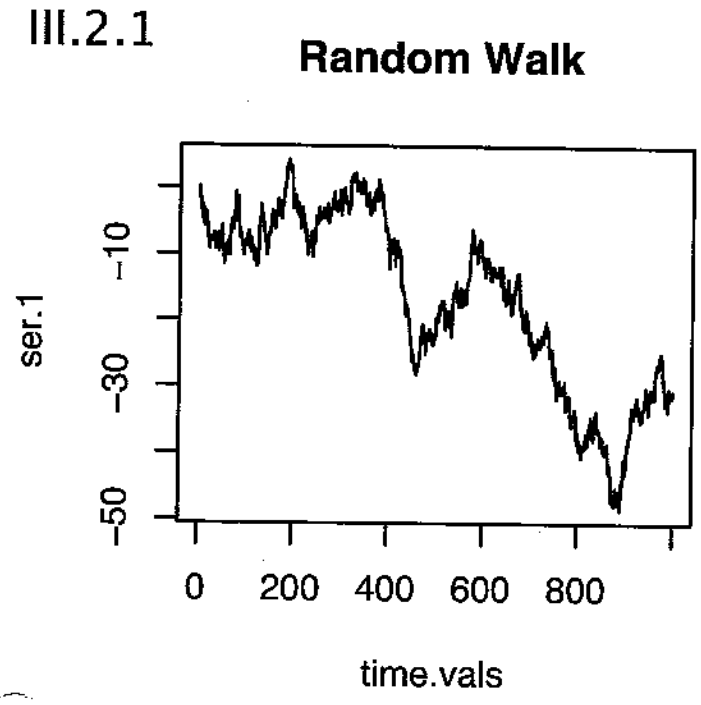
```
for (rep.lisa in 0:4){
for (iter in 1:100)
{
cen.var[1:iter,(iter+(n.samples*rep.lisa))] <-
self.censor(walk.var[1:iter,(iter+(n.samples*rep.lisa))])
}}
#Creates a matrix that uses observations from 1 to 100
#And extracts the self-censored count.

num.walk <-apply(cen.var,2,counter.fun)
lisa <-cbind(rep(1:100,5),num.walk)
colnames(lisa) <-c("Observations","Count")

lisa <-as.data.frame(lisa)
lisa.mod <-lm(Count~Observations,data=lisa)
summary(lisa.mod)
lisa.log.mod <-lm(log(Count)~Observations,data=lisa)
summary(lisa.log.mod)

pdf("/home/gecko/MSUDocs/GLMProj/GLMProjSum/Count.Orig.pdf")
plot(lisa,main="Random Walk Data, Counts vs Original Observations",cex=.2)
lines(ksmooth(lisa[,1],lisa[,2],bandwidth=5),col=3,lty=7)
abline(lisa.mod,col=2,lty=5)
dev.off()
```

K-S.R.txt

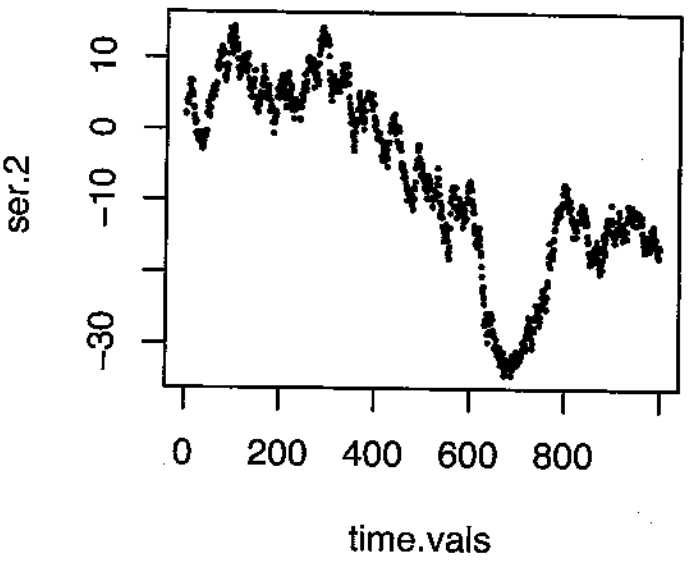
```
#Cody L. Custis
#K-S.R.Win.txt
#Used To Get Graphs With Significance And Power
source("/home/gecko/MSUDocs/GLMProj/GLMProjSum/useful.R.txt")
#Loads stand.unif,self.censor,random.walk
set.seed(112447)
n.samples <-200
l.samples <-1,000

unif.var=matrix(c(1),nrow=l.samples,ncol=n.samples)
for (iter in 1:n.samples)
{
unif.var[,iter] <-runif(l.samples)
}

exp.var=matrix(c(1),nrow=l.samples,ncol=n.samples)
for (iter in 1:n.samples)
{
exp.var[,iter] <-rexp(l.samples)
}

norm.var=matrix(c(1),nrow=l.samples,ncol=n.samples)
for (iter in 1:n.samples)
{
norm.var[,iter] <-rnorm(l.samples)
}

walk.var=matrix(c(1),nrow=l.samples,ncol=n.samples)
for (iter in 1:n.samples)
{
trend <-rep(0,l.samples)
noise <-rnorm(l.samples)
walk.var[,iter] <-random.walk(trend,noise)
}

#Be patient, may take a long time.
censored.exp <-apply(exp.var,2,self.censor)
censored.unif <-apply(unif.var,2,self.censor)
censored.norm <-apply(norm.var,2,self.censor)
censored.walk <-apply(walk.var,2,self.censor)
```

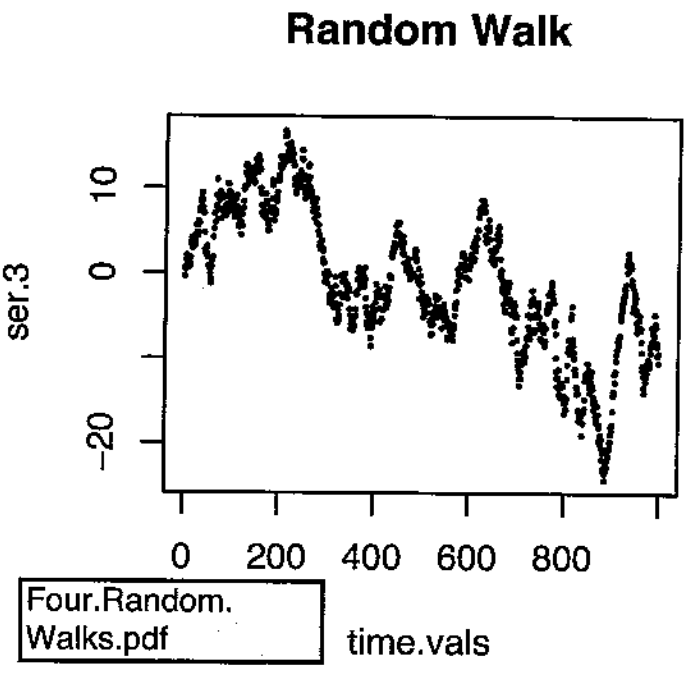
```
stand.exp <-apply(censored.exp,2,stand.unif)
stand.unif.mat <-apply(censored.unif,2,stand.unif)
stand.norm <-apply(censored.norm,2,stand.unif)
stand.walk <-apply(censored.walk,2,stand.unif)


array.exp <-apply(stand.exp,2,extract.fun)
array.unif <-apply(stand.unif.mat,2,extract.fun)
array.norm <-apply(stand.norm,2,extract.fun)
array.walk <-apply(stand.walk,2,extract.fun)



#Efficient way to test two distributions
exp.exp.mat <-c(1:1,000)
for (iter in 1:1,000)
{
samp.1 <-sample(1:200,1)
samp.2 <-sample(1:200,1)
exp.exp.mat[iter] <-
ks.test(jitter(array.exp[[samp.1]]),array.walk[[samp.2]])$p.value
}
cuts <-(1:100)/100
rejects.exp.exp<-cuts
for (iter in 1:100)
{
rejects.exp.exp[iter] <-length(exp.exp.mat[exp.exp.mat<cuts[iter]])
}
power.exp.exp <-rejects.exp.exp/1,000

postscript("/home/gecko/MSUDocs/GLMProj/GLMProjSum/goofing.ps")
plot(cuts <-(1:100)/100,power.exp.exp,xlab="Significance",ylab="Power",main="Walk
vs. Unif",ylim=c(0,1));abline(a=0,b=1)
dev.off()
```

III.2.1

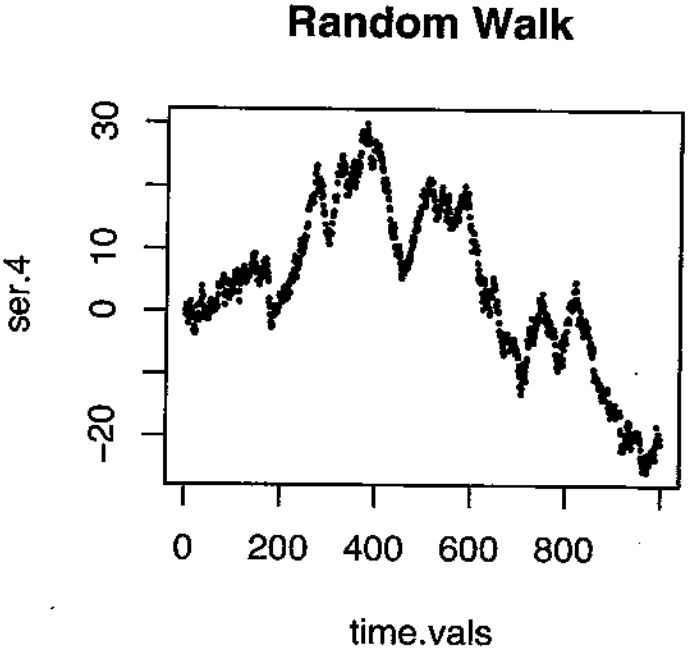


Random Walks Generated Using R Function

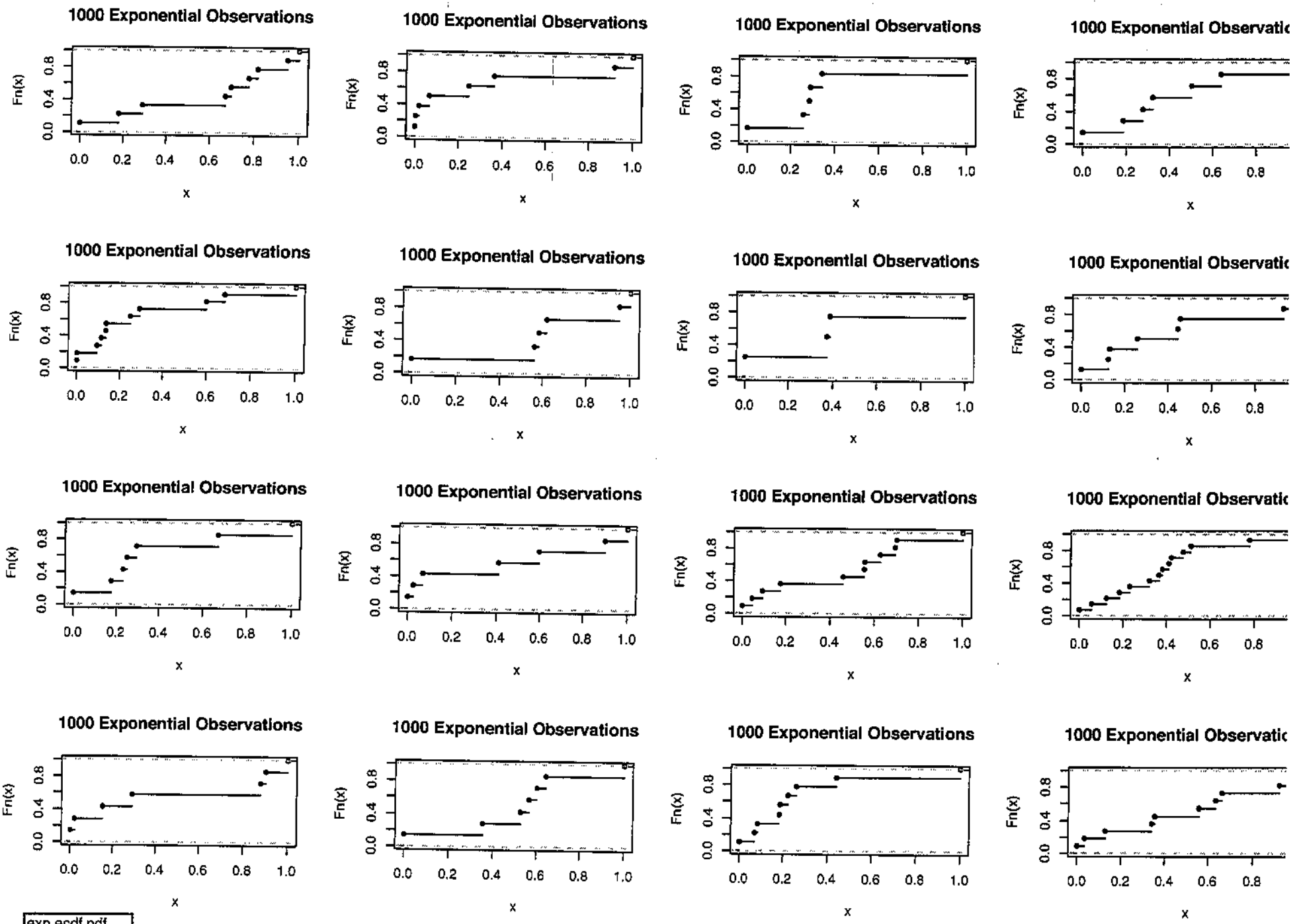




Four.Random.
Walks.pdf

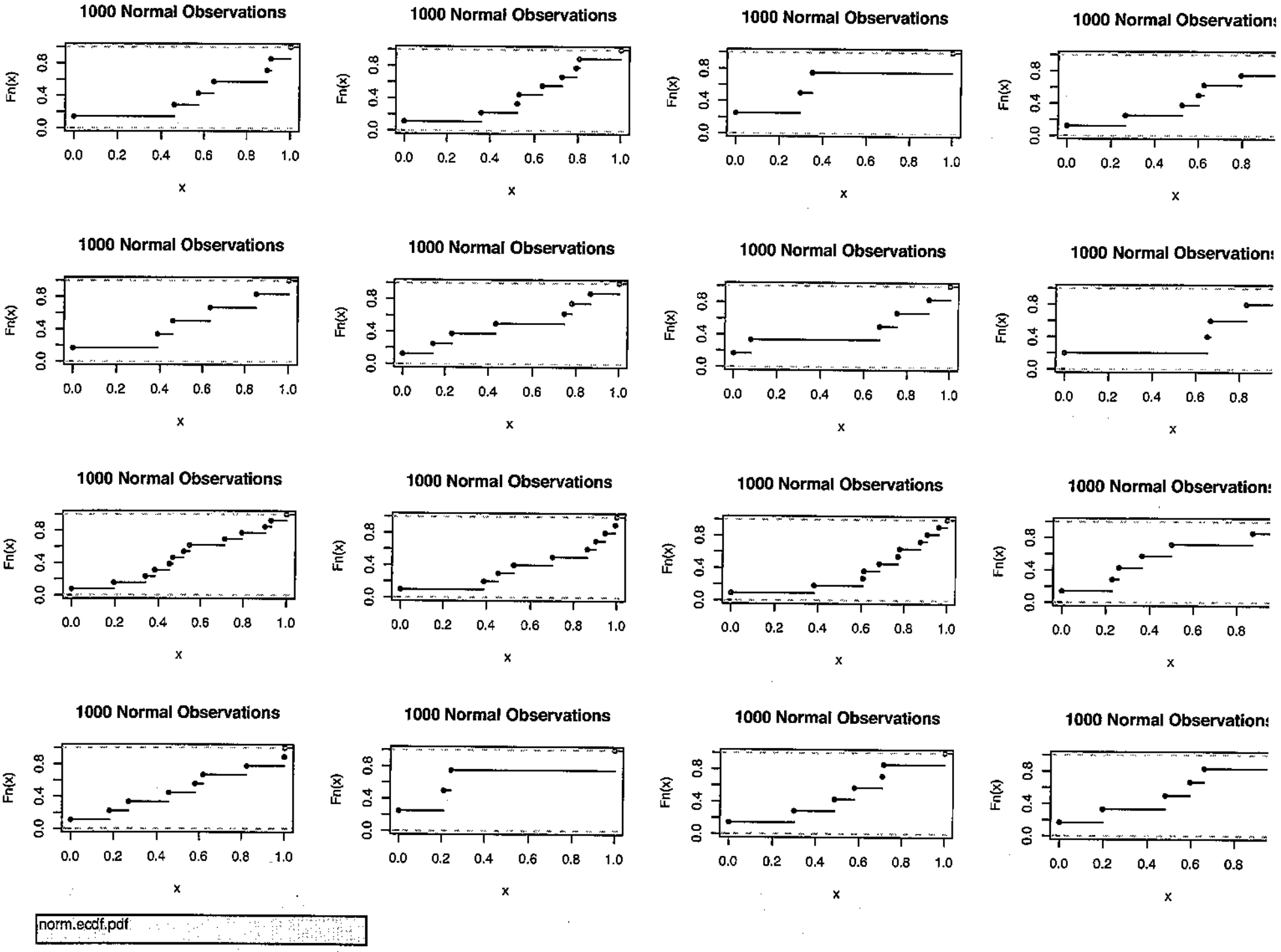


III.5 Empirical c.d.f.s after self-censoring for exponential distribution.



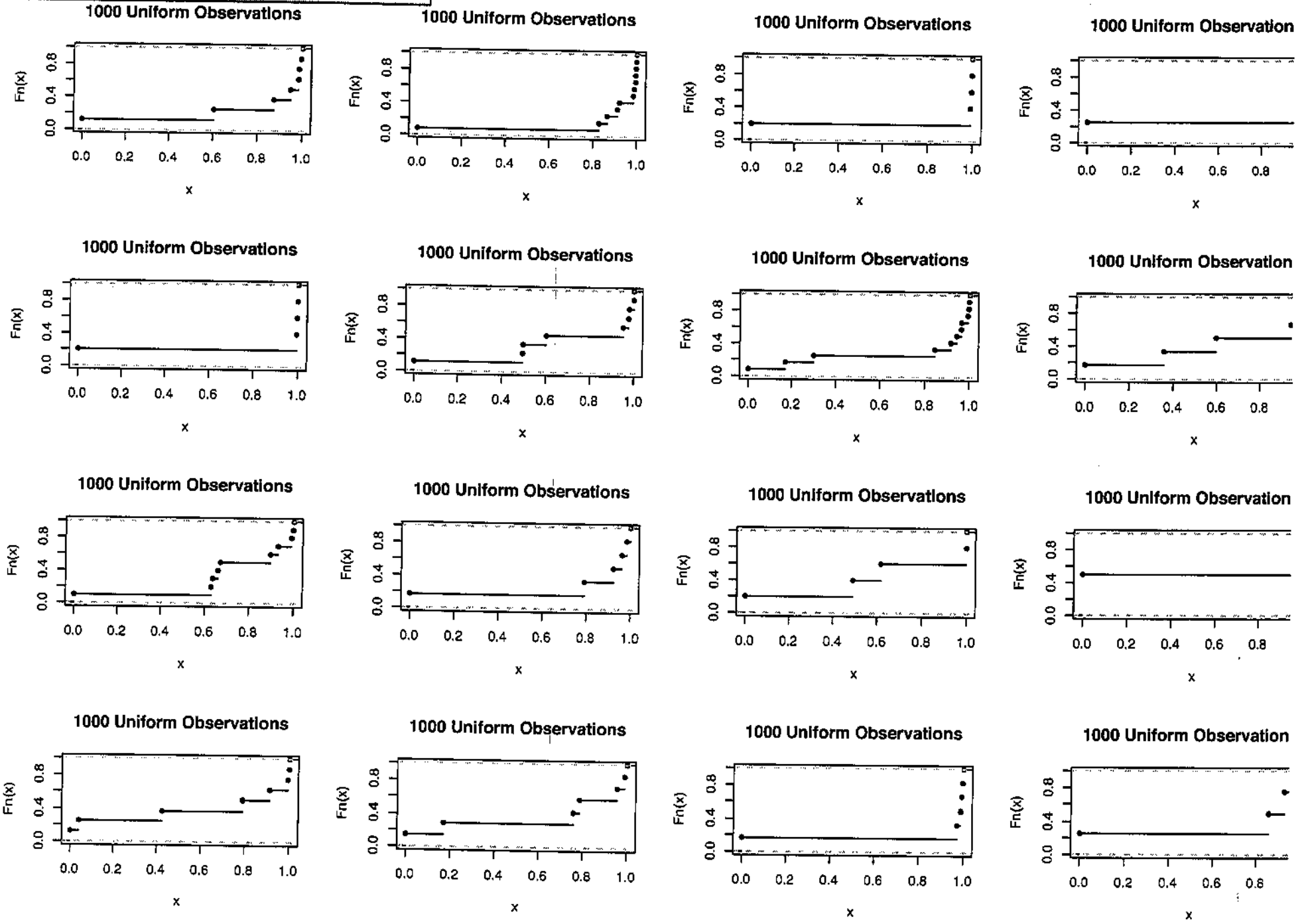
exp.ecdf.pdf

III.5 Empirical c.d.f.s after self-censoring for normal distribution.



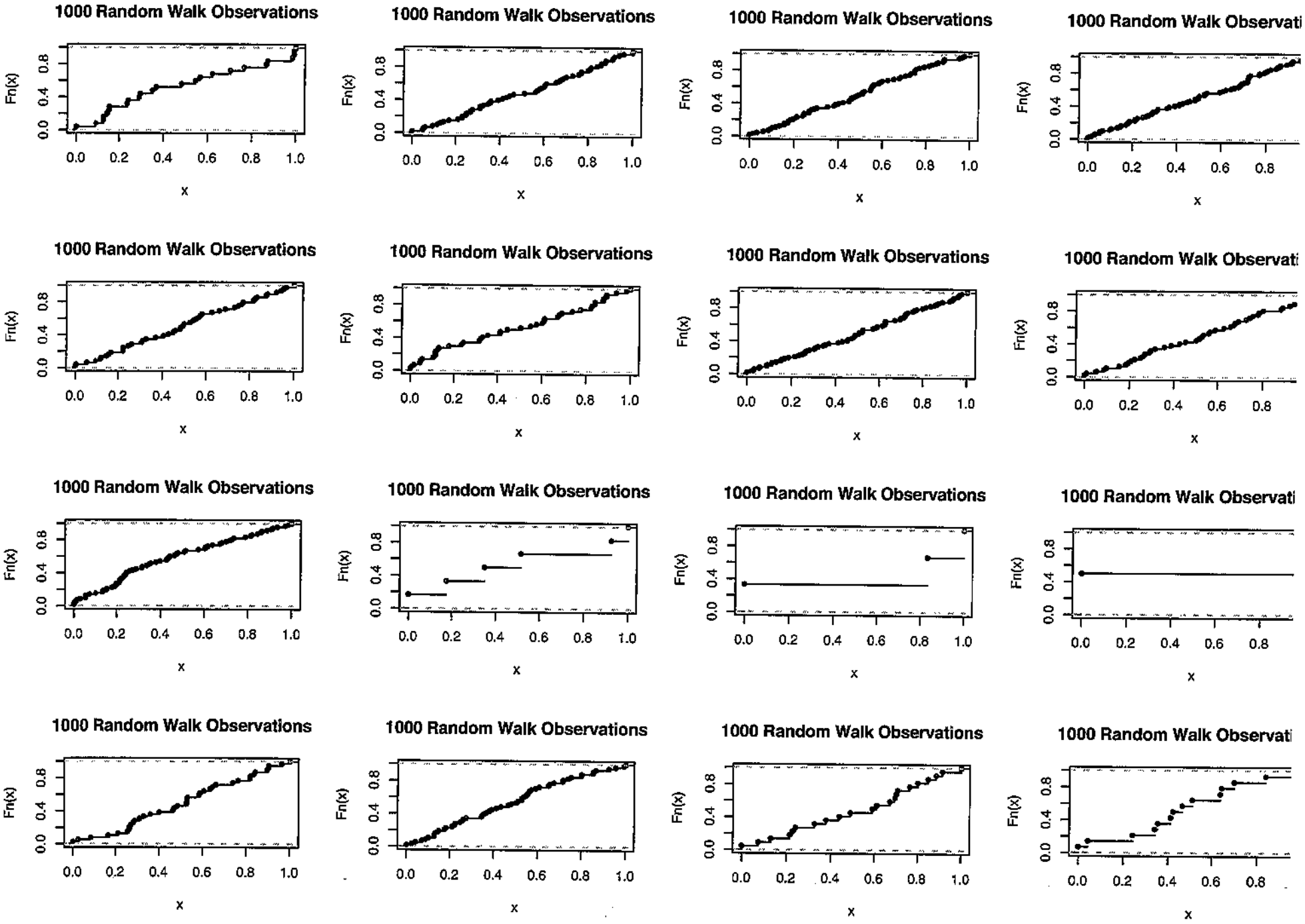
norm.ecdf.pdf

III.5 Empirical c.d.f.s after self-censoring for uniform distribution.



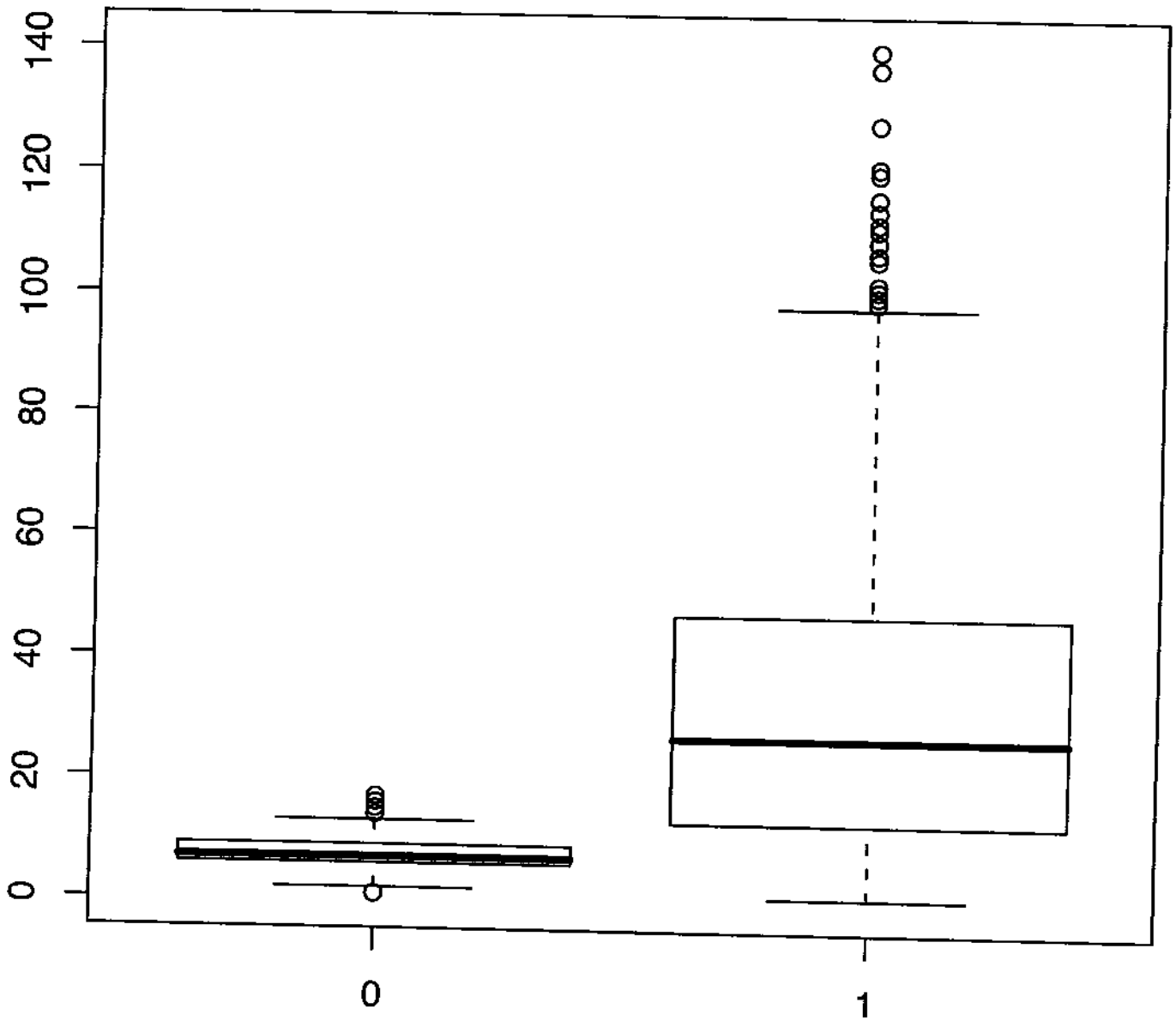
unif.ecdf.pdf

III.5 Empirical c.d.f.s after self-censoring for random walk distribution.



walk.ecdf.pdf

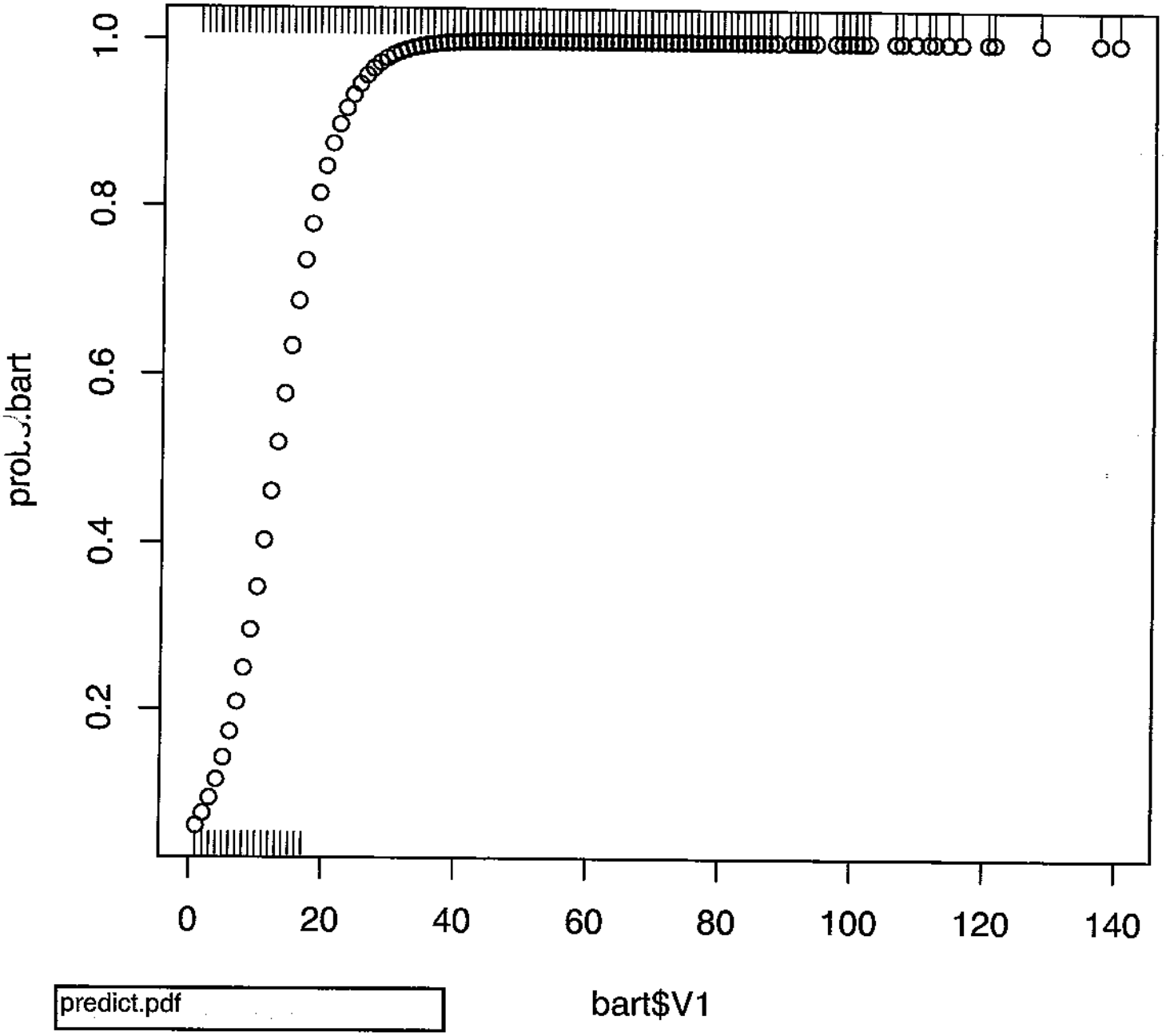
III.6 Boxplots Comparing Counts After Self-Censoring From Exponetial and Random Walk Distributions



comp.boxplots.pdf

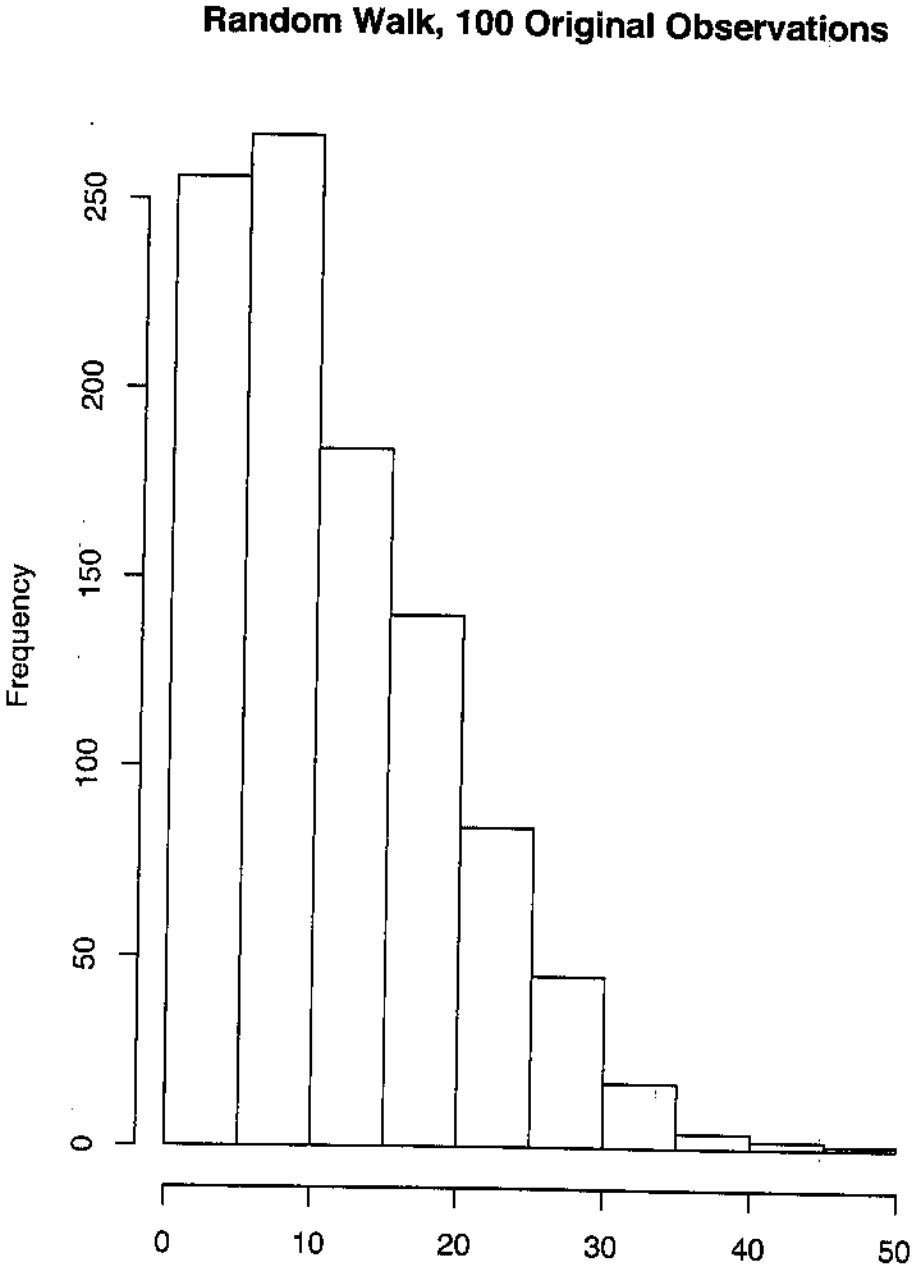
Exp=0
Walk=1

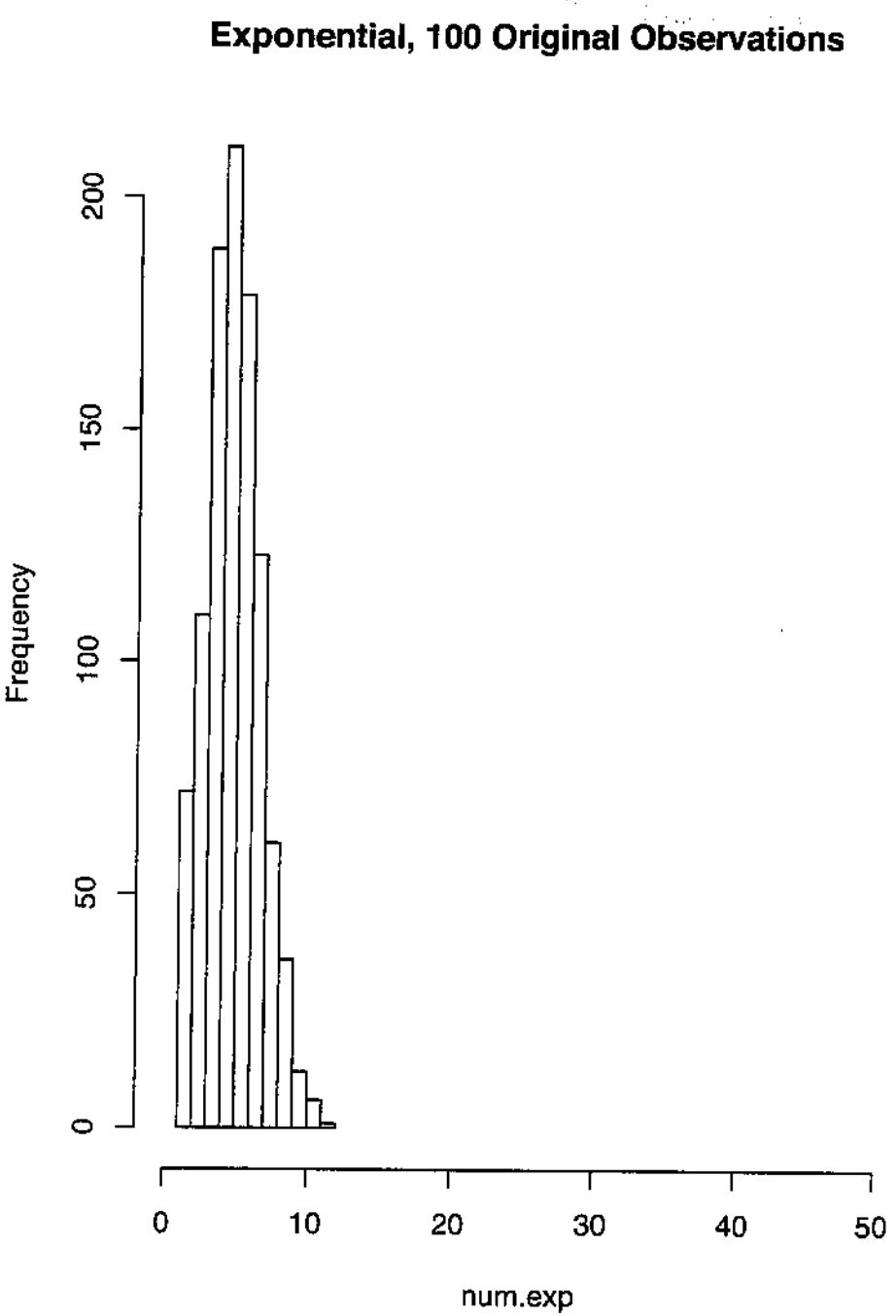
III.6 Plot of logistic regression curve, with observed values fitted as a 'rug.'



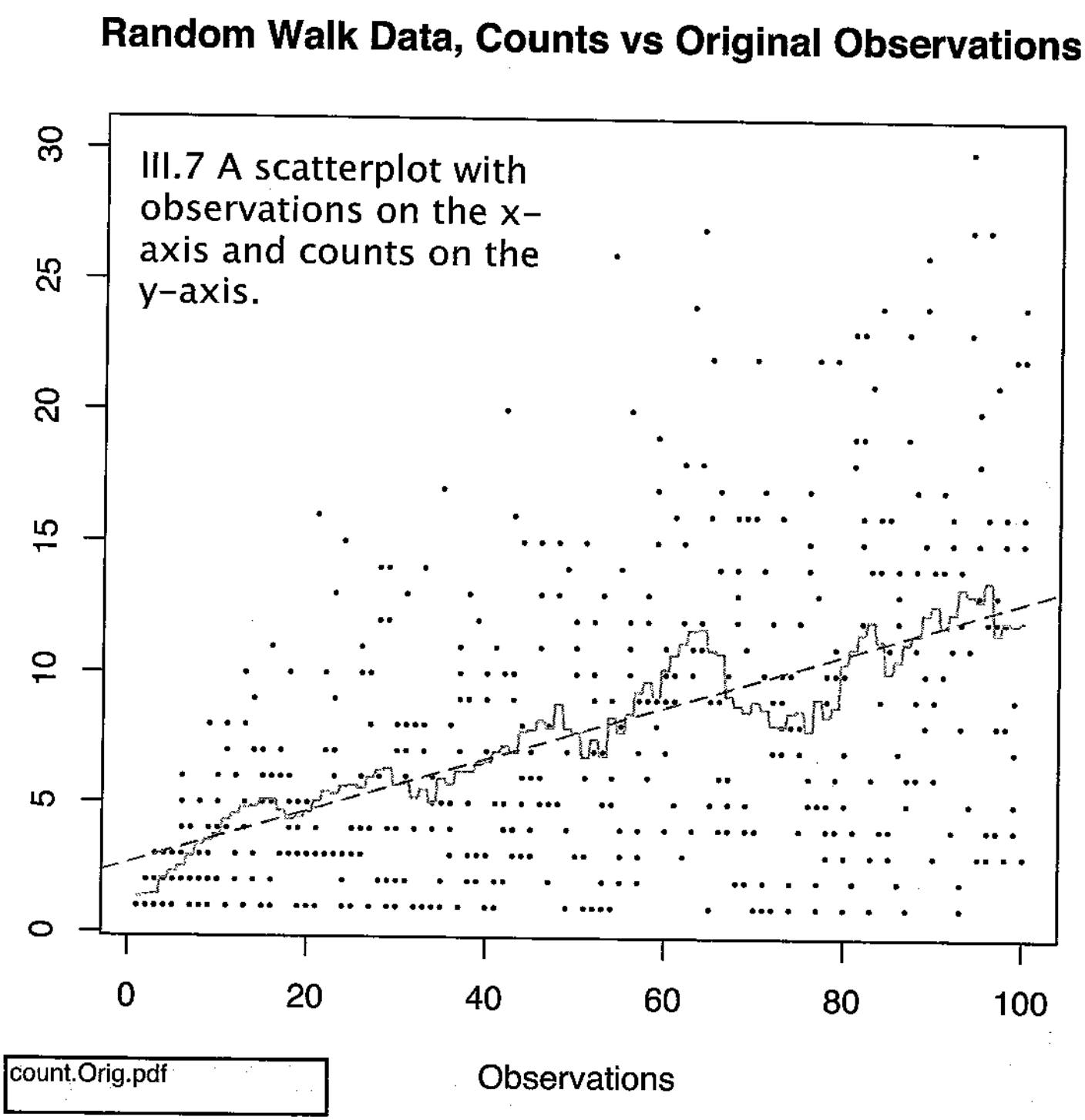
predict.pdf

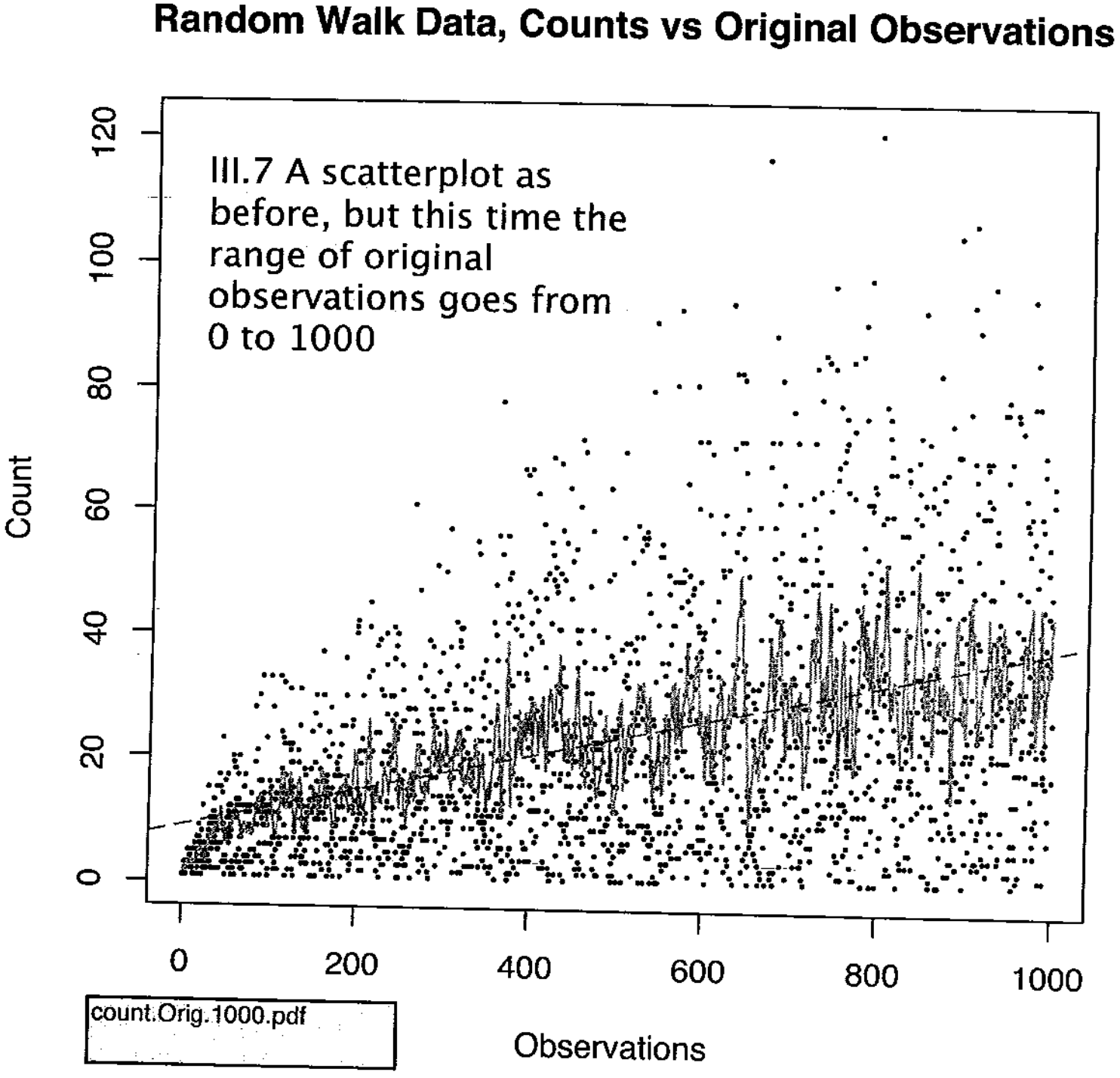
III.6 Histograms of counts after self-censoring. Only 100 original observations were used.



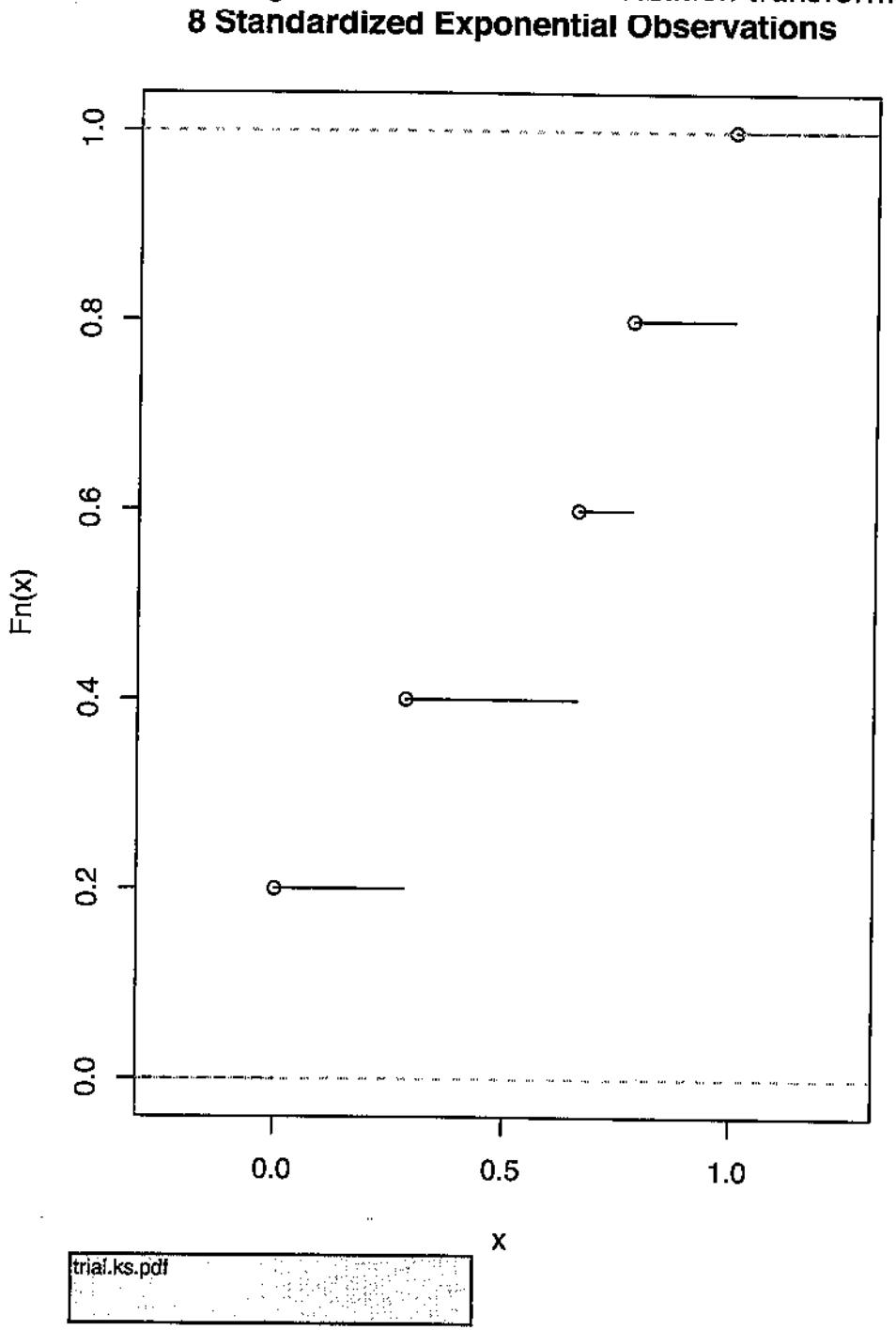


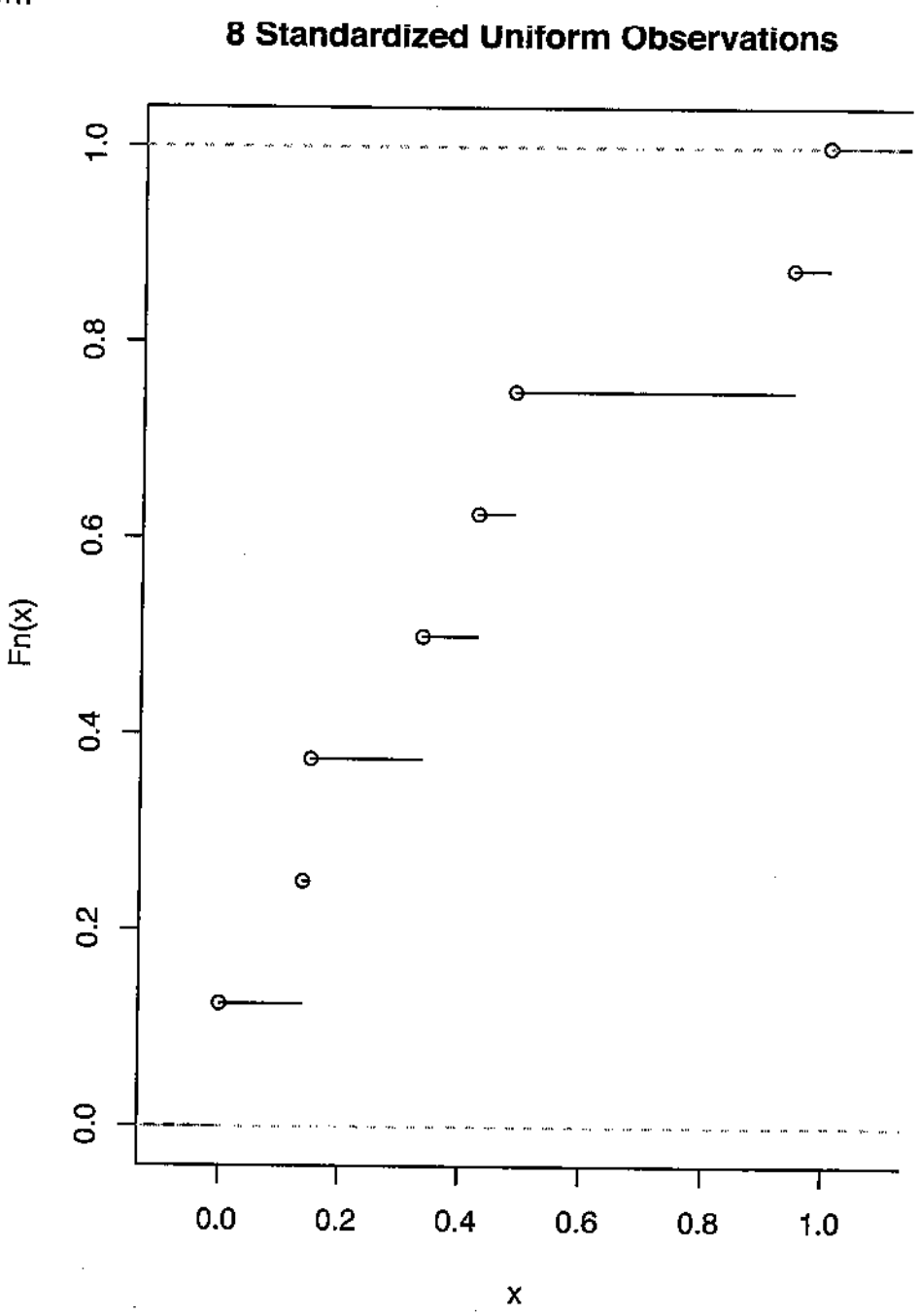
hist.walk.exp.100.pdf


Random Walk Data, Counts vs Original Observations
III.7 A scatterplot with observations on the x-axis and counts on the y-axis.
Count
0
5
10
15
20
25
30
0
20
40
60
80
100
Observations
count.Orig.pdf

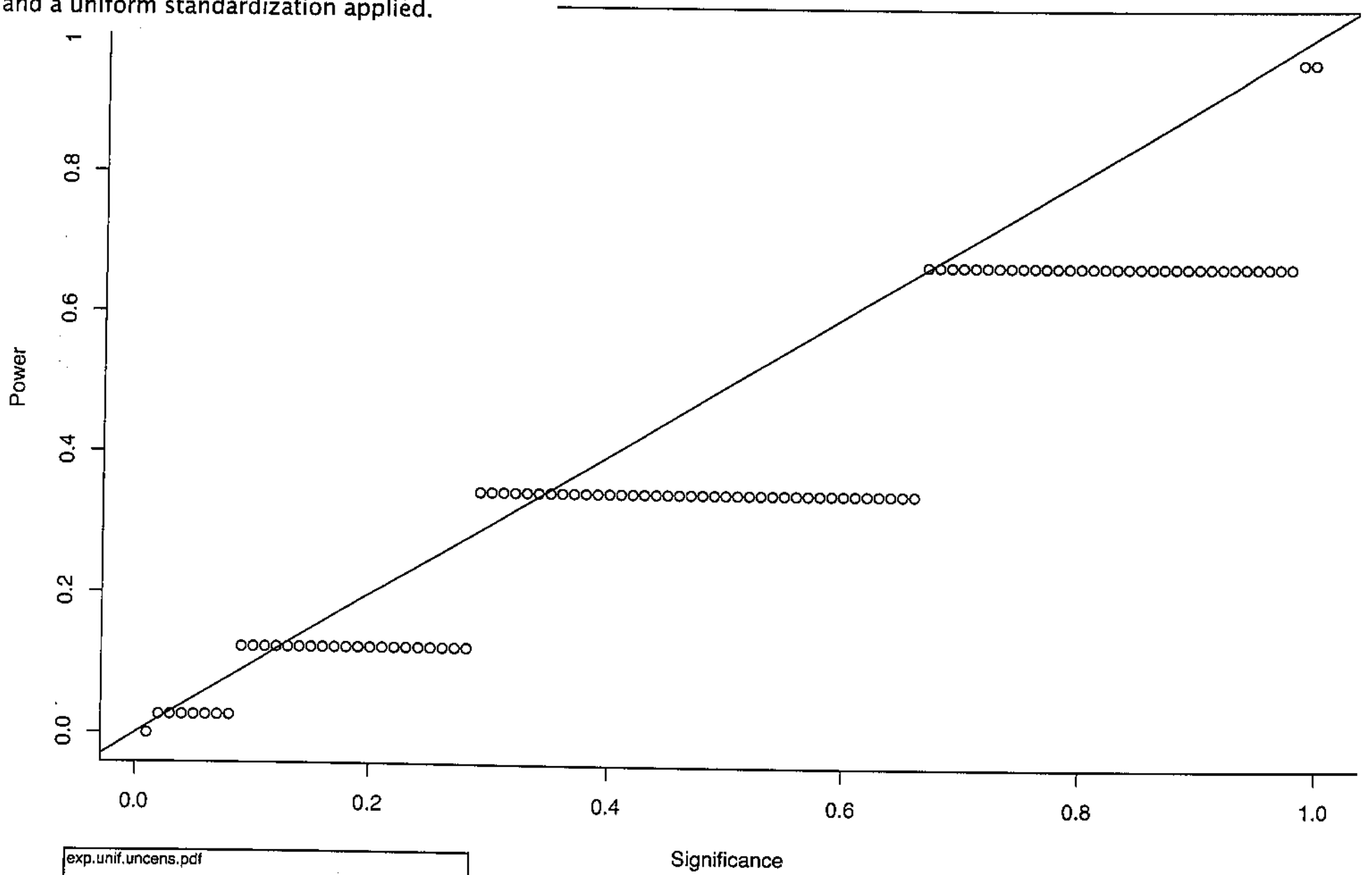

Random Walk Data, Counts vs Original Observations
III.7 A scatterplot as before, but this time the range of original observations goes from 0 to 1000
Count
0
20
40
60
80
100
120
0
200
400
600
800
1000
Observations
count.Orig.1000.pdf

III.9 Empirical cumulative distribution functions for 8 observations from the uniform and exponential distributions, which have undergone a uniform standardization transformation.



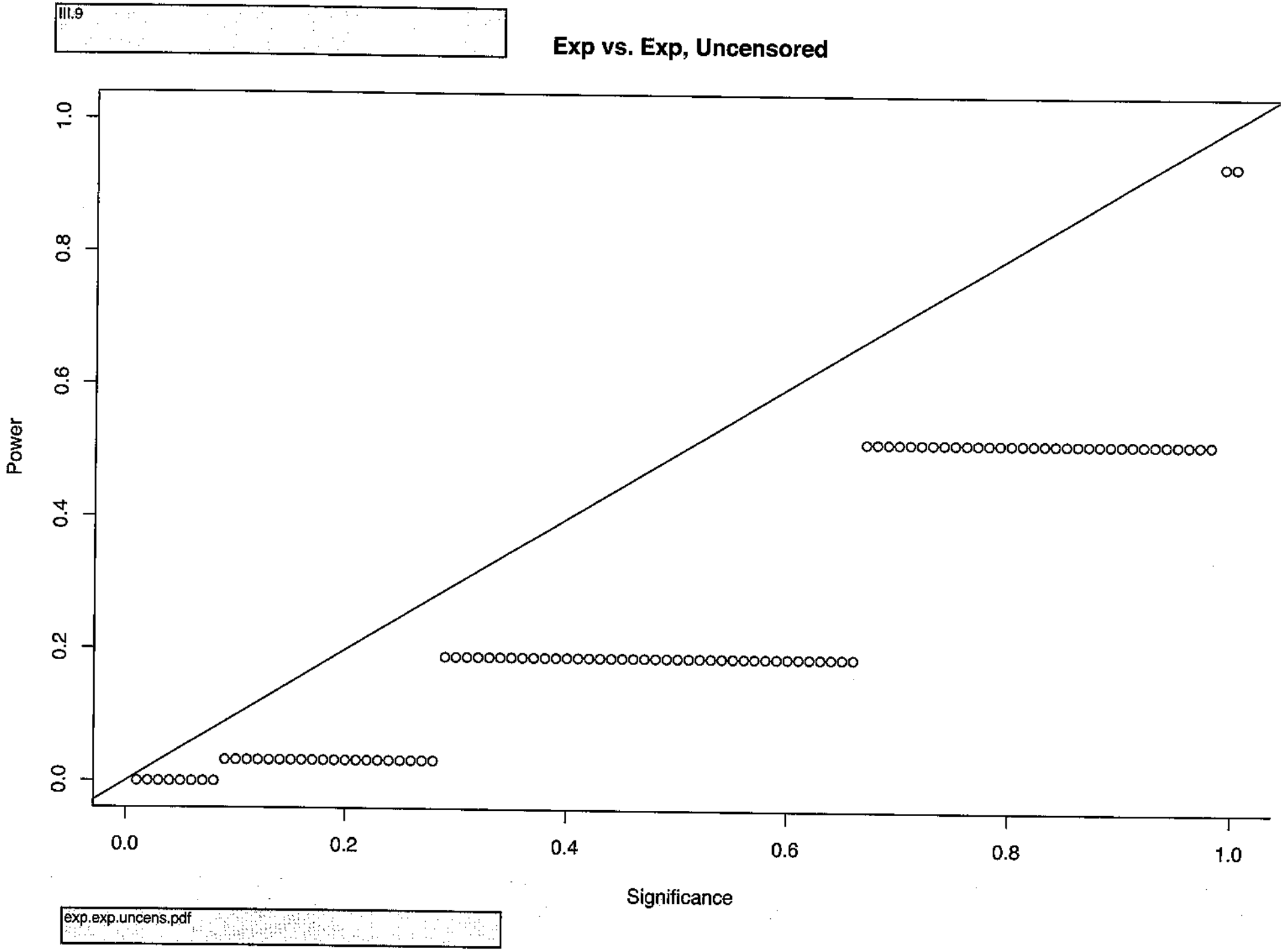


III.9 This plot compares the power of the K–S test to declare that the population c.d.f.s are not the same versus the significance level chosen, with 8 observations generated from each distribution, and a uniform standardization applied.



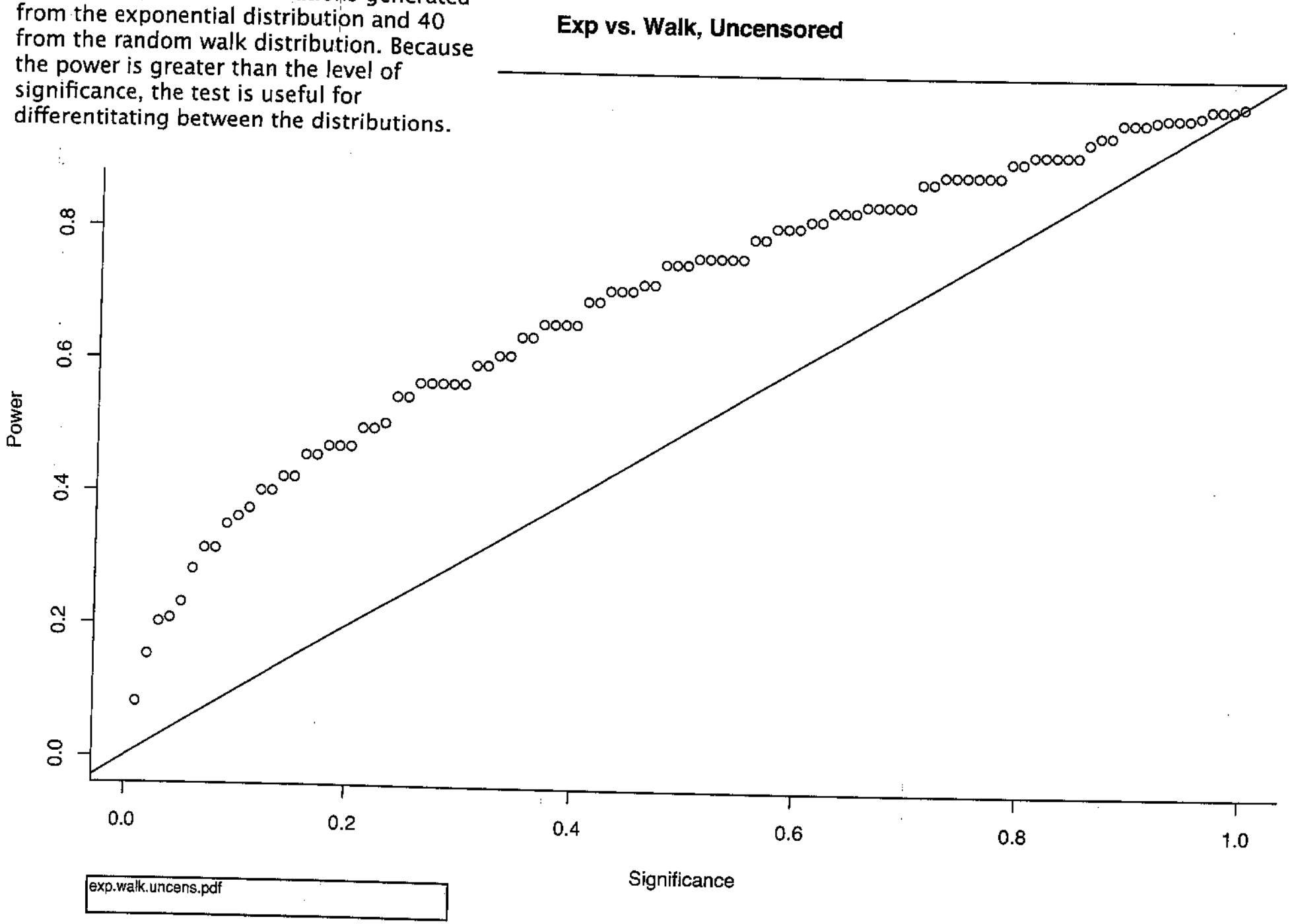
exp.unif.uncens.pdf

III.9



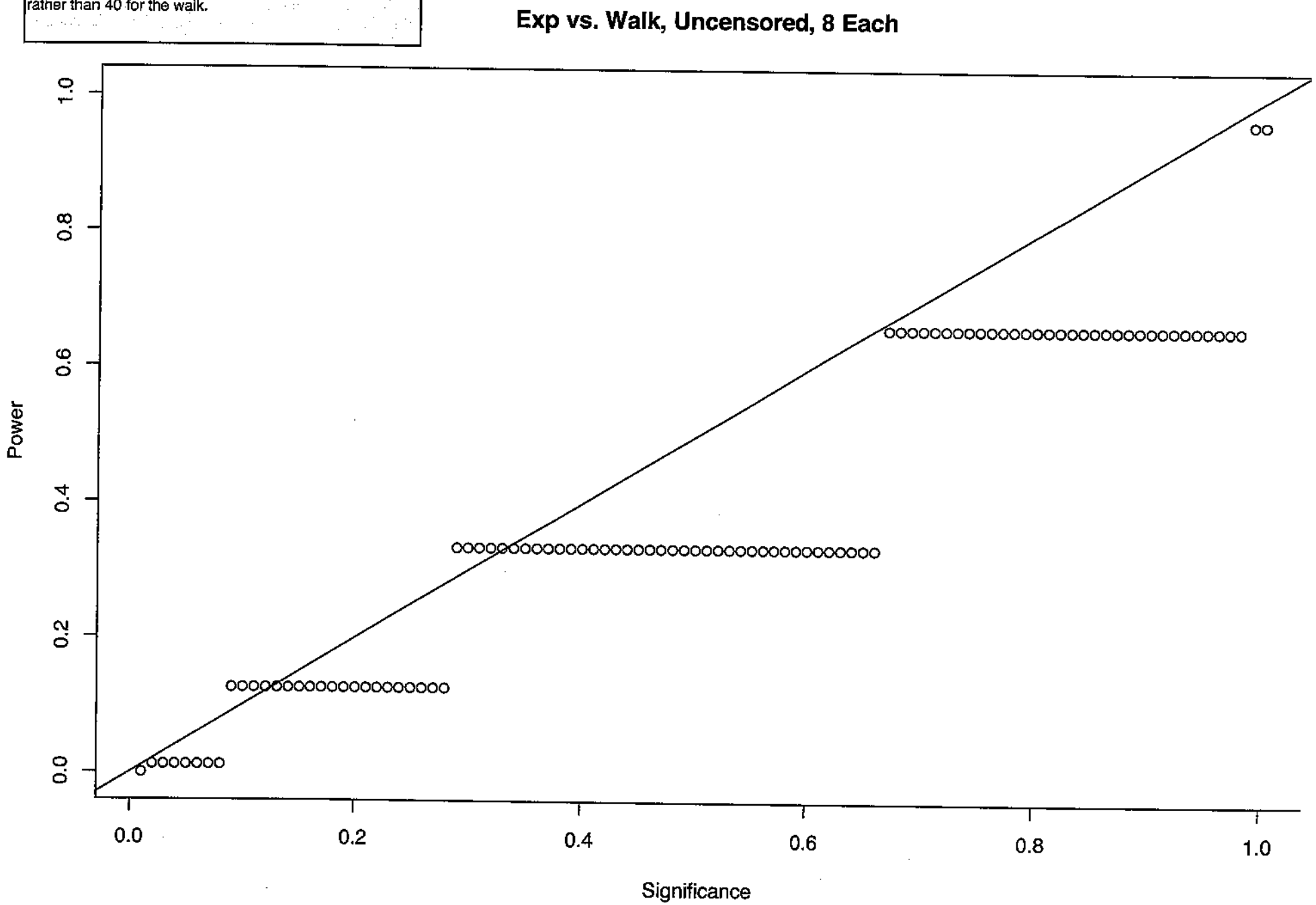
exp.exp.uncens.pdf

III.9 As before, this compares power and significance, with 8 observations generated from the exponential distribution and 40 from the random walk distribution. Because the power is greater than the level of significance, the test is useful for differentitating between the distributions.



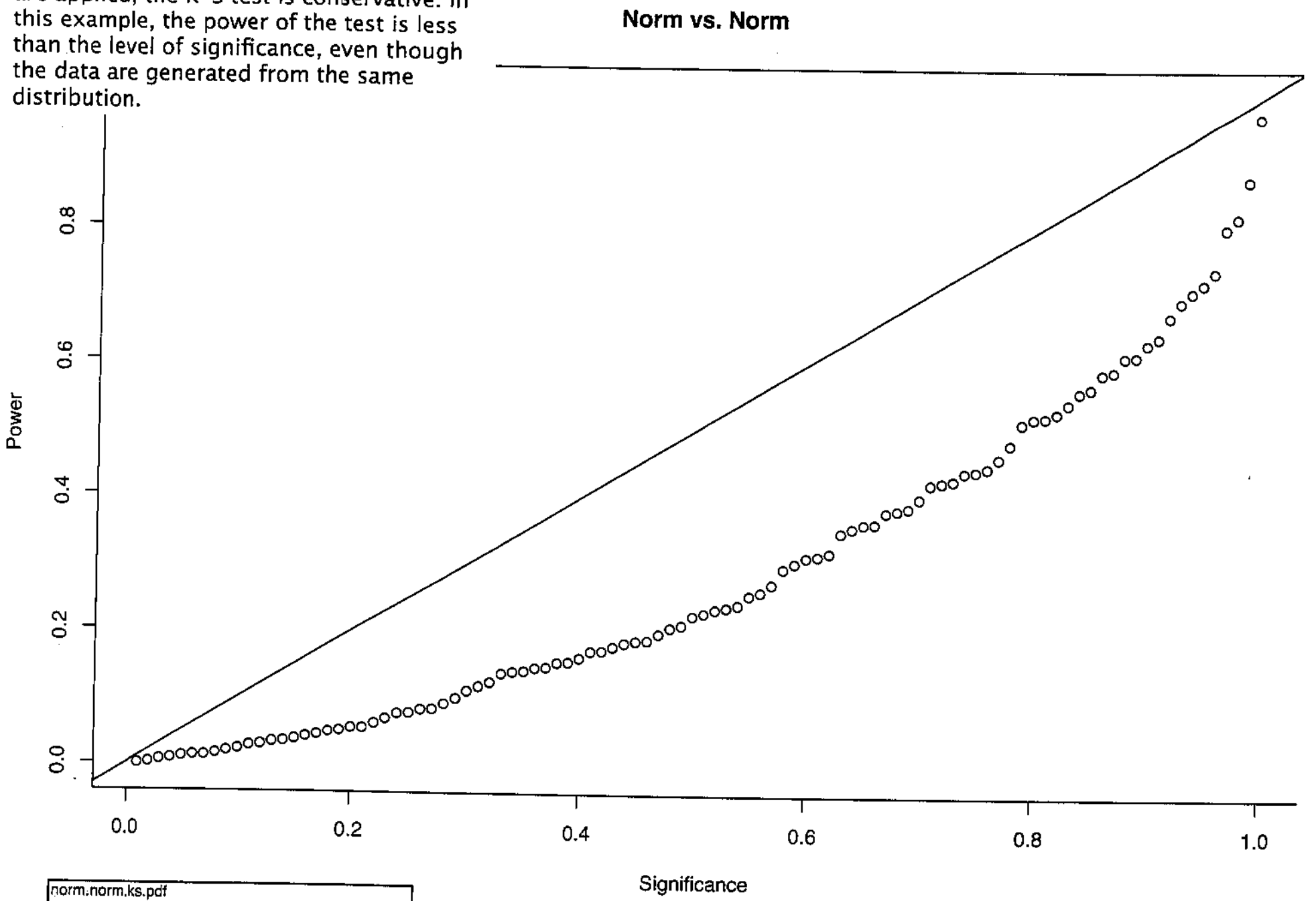
exp.walk.uncens.pdf

III.9 Generated as before, but it only uses 8 observations from both exponential and random walk distributions, rather than 40 for the walk.



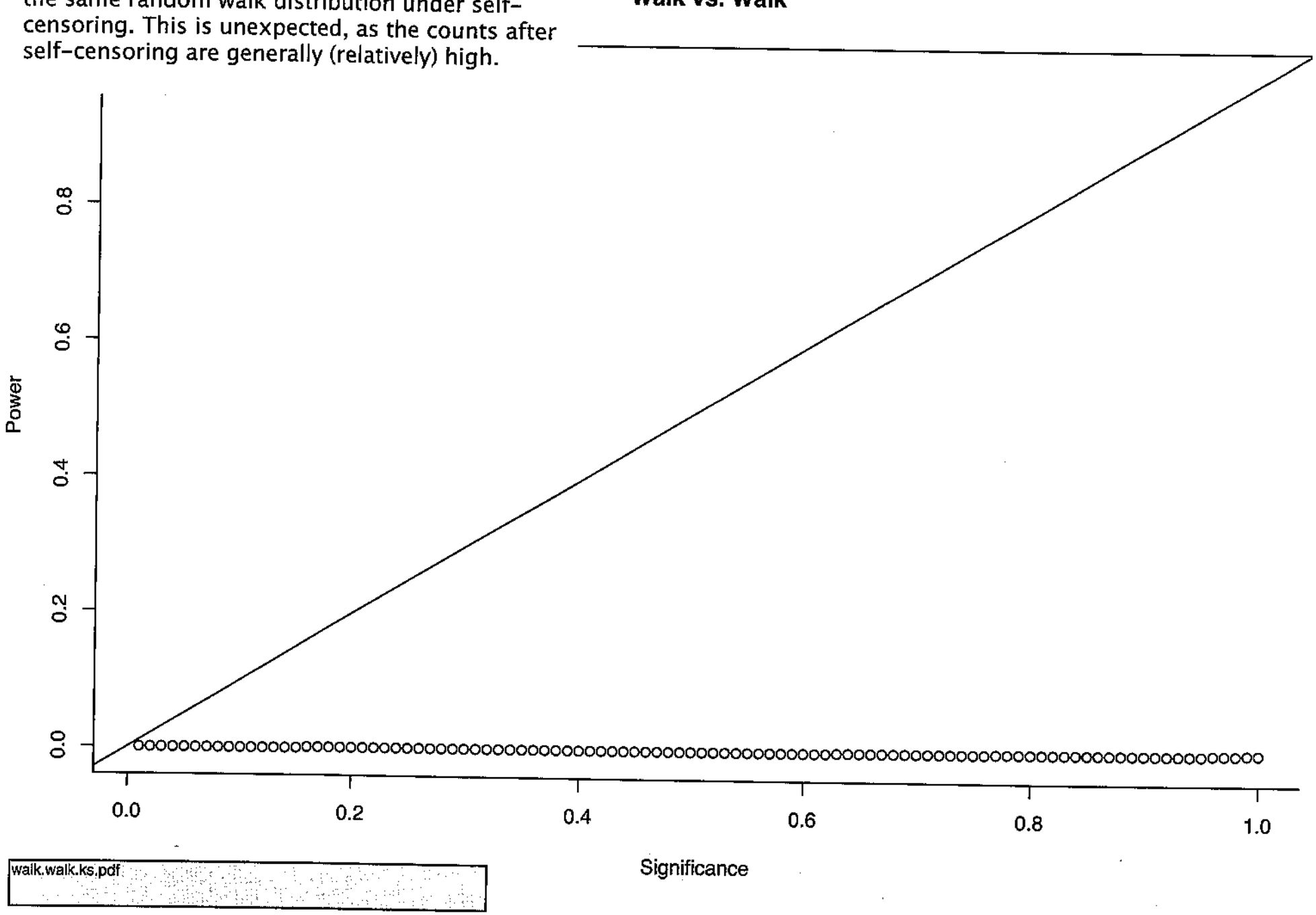
exp.walk.uncens.small.pdf

III.10 When self-censoring transformations are applied, the K-S test is conservative. In this example, the power of the test is less than the level of significance, even though the data are generated from the same distribution.



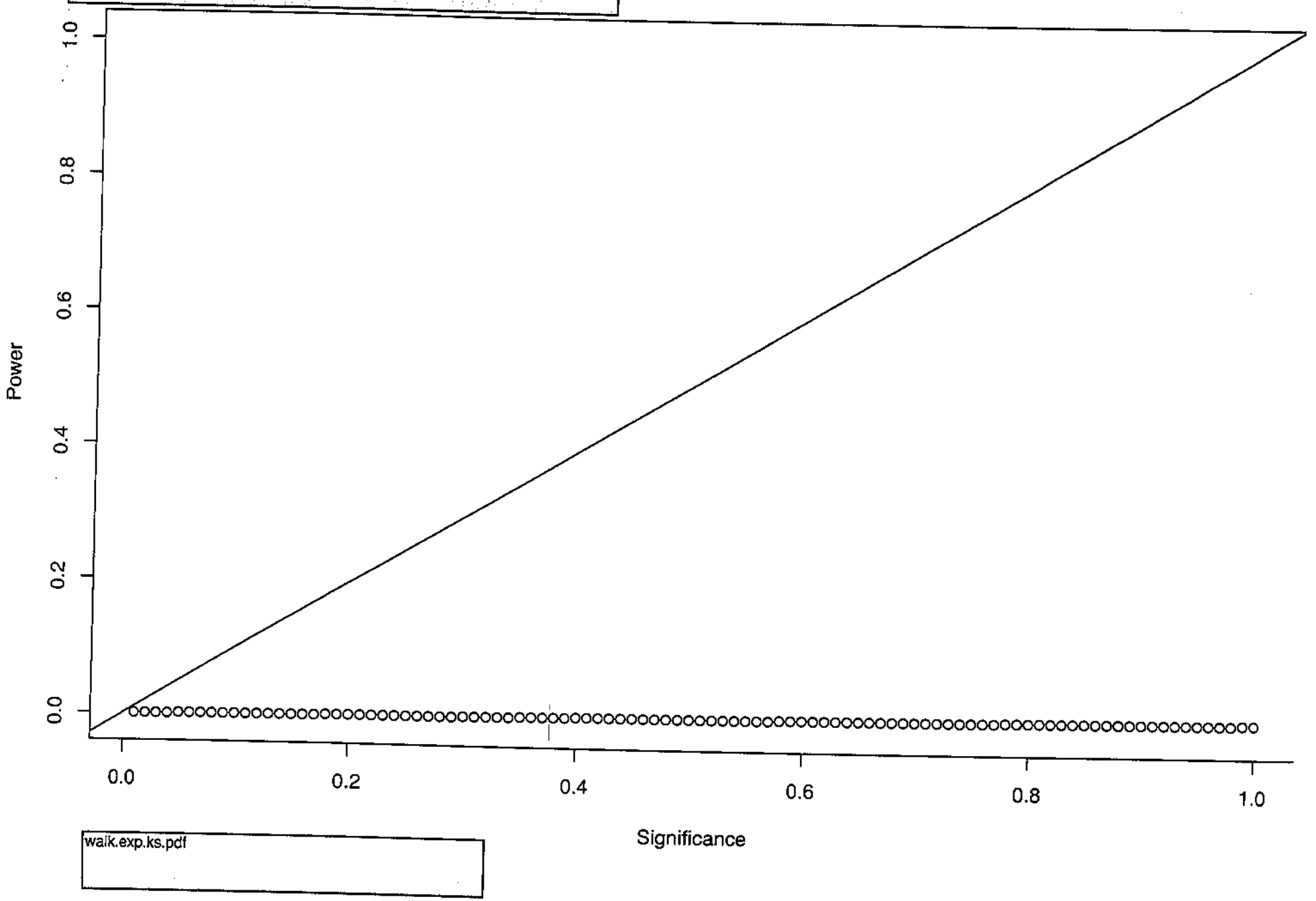
norm.norm.ks.pdf

III.10 The K-S test has high p-values under simulation when comparing data generated from the same random walk distribution under self-censoring. This is unexpected, as the counts after self-censoring are generally (relatively) high.

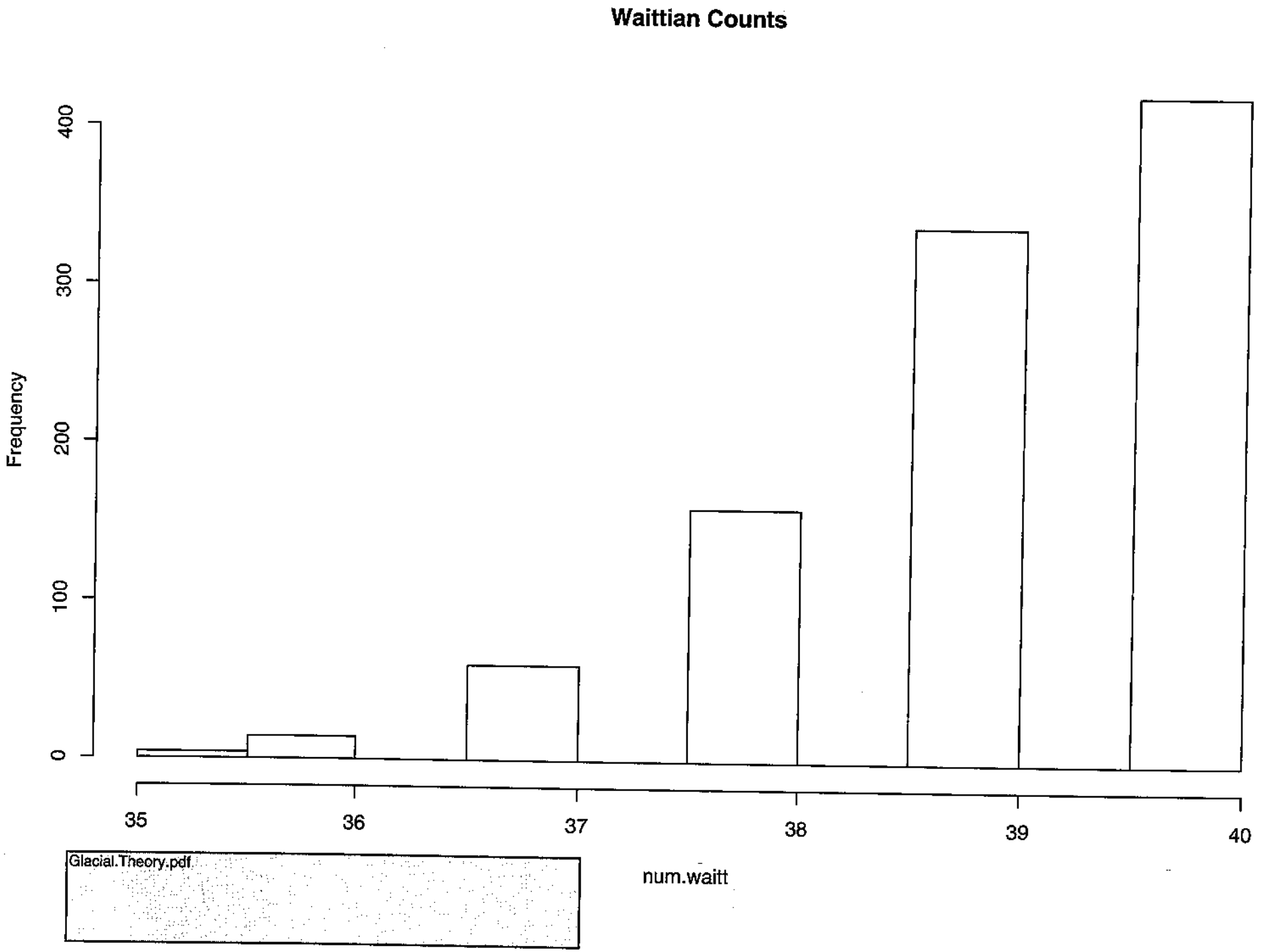


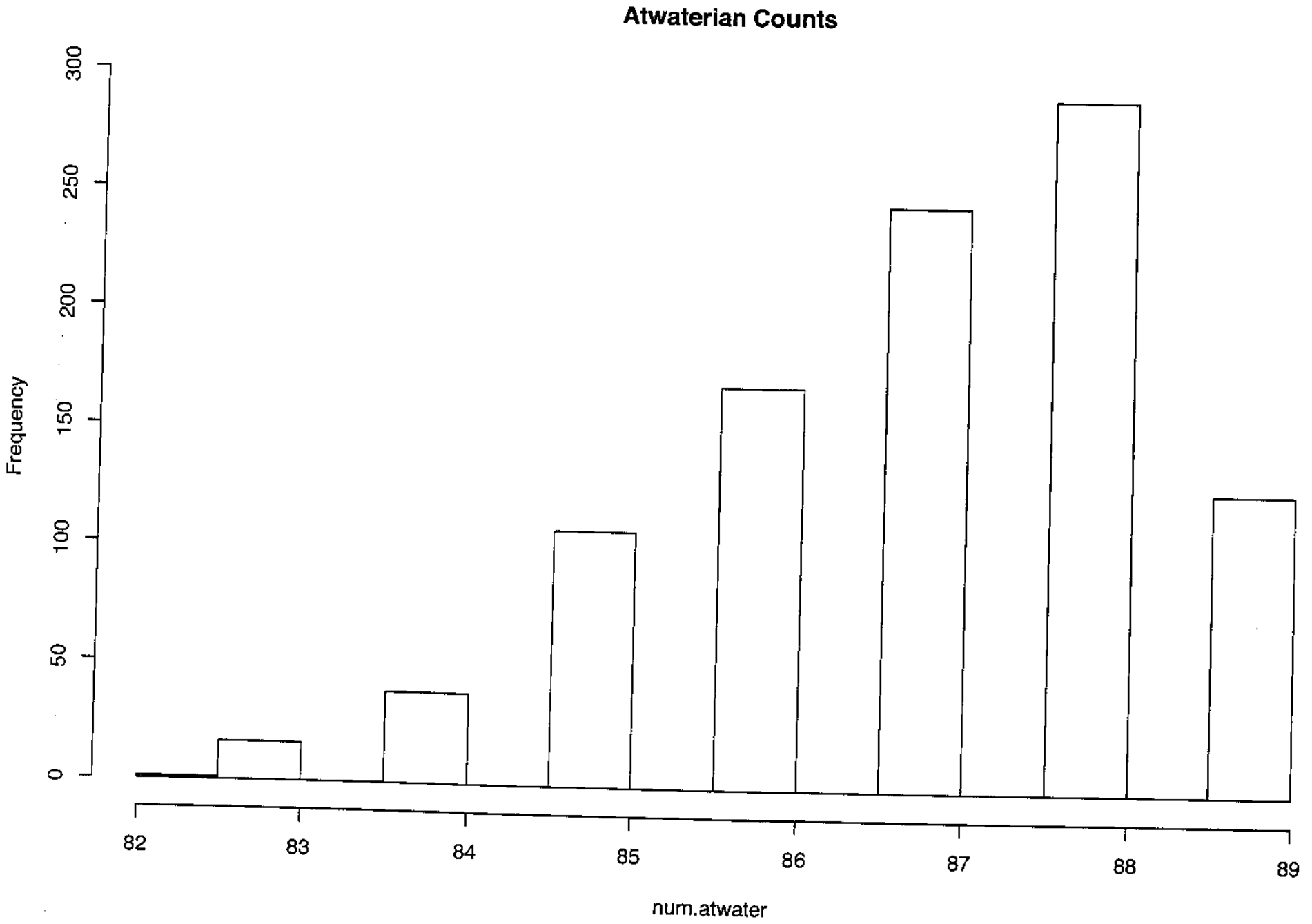
walk.walk.ks.pdf

III.10 Now that self-censoring is applied, the K-S test lacks the ability to differentiate between the random walk and exponential distributions, at sample sizes comparable to those used without censoring.



walk.exp.ks.pdf


Waittian Counts
Frequency
0
100
200
300
400
35
36
37
38
39
40
num.waitt
Glacial.Theory.pdf


Atwaterian Counts
Frequency
0
50
100
150
200
250
300
82
83
84
85
86
87
88
89
num.atwater

Appendix 4: Bibliography

"An Introduction to R" Version 2.3.1 (2006-06-01)

Alt, David *Glacial Lake Missoula And Its Humongous Floods* Mountain Press Publishing Company Missoula, MT 2002

Atwater, Brian F. Periodic floods from glacial Lake Missoula into the Sanpoil arm of glacial Lake Columbia, northeastern Washington Geology v. 12 p 464-467, 1984
- Another correlation between varves and the Missoula floods.

Baker, Victor R. Paleohydrology and Sedimentation of Lake Missoula Flooding in Eastern Washington. Geological Society of America Special Paper 144 1973
- Gives useful estimates of the discharge in the Missoula floods.

Baker, Victor R. The Spokane Flood Controversy and the Martian Outflow Channels *Science*, New Series, Vol. 202, No. 4374 (Dec. 22, 1978), 1249-1256
- Contains a good overview of the controversy behind the Spokane Flood controversy.

Baker, Victor R. Surprise Endings to Catastrophism and Controversy on the Columbia GSA Today. Vol. 5, No. 9 September 1995
- Provides another overview of the development of Glacial Lake Missoula as a source for the channeled scabland.

Baker, V. R., and Bunker, R.C., 1985, Cataclysmic late Pleistocene flooding from glacial Lake Missoula: A review: Quaternary Science Reviews, v. 4, p. 1–41.
- Gives another overview, and criticizes the Waittian view. Relies on volume of water and flood depth.

Bretz, J Harlen The Channeled Scablands Of The Columbia Plateau Journal of Geology 31, 617-649 (1923)
- Suggests flooding as a source for channeled scablands of Eastern Washington.

Bretz, J Harlen The Dalles Type Of River Channel Journal of Geology 32, 139-149 (1924)

Bretz, J Harlen The Channeled Scabland Of Eastern Washington Geographical Review, Vol. 19, No. 3 (Jul., 1928), 446-477
-Contains information on the channeled scabland of Eastern Washington.

Bretz, J Harlen Lake Missoula And The Spokane Flood Geological Society Of America Bulletin 41 pp 92-93, 1930

Casella, George and Roger L. Berger *Statistical Inference* First Edition Duxbury 1990

Conover, W.J. *Practical Nonparametric Statistics* Third Edition Wiley 1999

Daniel, Wayne W. *Applied Nonparametric Statistics* Second Edition Duxbury 1990

DeGroot, Morris H. and Mark J. Schervish *Probability and Statistics* 3rd Edition Addison and Wesley 2002

Harding, H.T. Possible Water Supply for the Creation of Channeled Scab Lands *Science*, New Series, Vol. 69, No. 1781 (Feb 15, 1929), 188-190
- Harding initially proposes the multi-flood hypothesis.

Hollander, Myles and Douglas A. Wolfe *Nonparametric Statisical Methods* Second Edition Wiley 1999

Montgomery, Douglas C. and Lynwood A. Johnson *Forecasting and Time Series Analysis* McGraw Hill 1976

Munro, Barbara Hazard *Statistical Methods for Health Care Research* 5th Edition Lippincott Williams & Wilkins 2005

O'Connor, James E. and Richard B. Waitt Beyond the Channeled Scabland: A field trip to Missoula flood features in the Columbia, Yakima, and Walla Walla valleys of Washington and Oregon - Part 1 Oregon Geology, Volume 57, Number 3, May 1995 p. 51-58

Pardee, J.T. The Glacial Lake Missoula Journal Of Geology, Vol. 18 p 376-386 1910

"R Language Definition Version 2.3.1" (2006-06-01)
http://cran.r-project.org/doc/manuals/R-lang.html#Functions

R Help Files Version 1.8.1 (2003-11-21)
- To the author's knowledge, this is the same help file found on current versions of R.

Shaw, John et. al. The Channeled Scabland: Back To Bretz? Geology: July 1999; v. 27; no. 7 p. 605-608

Smith, Gary A. Missoula flood dynamics and magnitudes inferred from sedimentology of slack-water deposits on the Columbia Plateau, Washington Geological Society of America Bulletin, v. 105, p. 77-100, January 1993

Waitt, Richard B. About Forty Last-Glacial Missoula Jökulhlaups Through Southern Washington. Journal of Geology 1980, vol. 88, p. 653-679
- Major paper proposing the multiple flood hypothesis.

Waitt, Richard B. Tens Of Successive, Colossal Missoula Floods At North And East Margins Of Channeled Scabland FRIENDS OF THE PLEISTOCENE, Rocky Mountain Cell Guidebook for 1983 Field Conference Day 2: August 1983

Waitt, Richard B. Case for periodic, colossal jökulhlaups from Pleistocene glacial Lake Missoula. Geological Society of America Bulletin, v. 96, p1271-1286, October 1985

Weisstein, Eric W. "High-Water Mark." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/High-WaterMark.html Last Modified June 22, 2004 Accessed August 2006

Weisstein, Eric W. "Running Maximum." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/RunningMaximum.html Last modified October 5, 2003 Accessed August 2006