LogNormal Distributions for Total Water Intake and Tap Water Intake by Pregnant and Lactating Women in the United States

David E. Burmaster
Alceon Corporation
PO Box 382669, Cambridge, MA 02238-2669
tel: 617-864-4300; fax: 617-864-9954
deb@Alceon.com

Abstract

Using probability plots and Maximum Likelihood Estimation (MLE), we fit LogNormal distributions to data compiled by Ershow et al. (1991) for daily intake of total water and tap water by 3 groups of women (controls, pregnant, and lactating; all between 15 - 49 years of age) in the United States. We also develop bivariate LogNormal distributions for the joint distribution of water ingestion and body weight for these 3 groups. Overall, we recommend the marginal distributions for water intake as fit by MLE for use in human health risk assessments.

Introduction and Data

In 1978, the US Department of Agriculture (USDA) conducted the Nationwide Food Consumption Survey (NFCS) to gather dietary information for 7 days on individuals living in randomly assigned nonmilitary households in the contiguous 48 states. (USDA, 1980). From the database for 30,770 persons who participated in the survey, the NFCS identified women in the age range from 15 - 49 years and then grouped them into 3 categories: (i) control women (nonpregnant and nonlactating; N = 6,201), (ii) pregnant women (N = 188); and (iii) lactating women (N = 77). Ershow et al. (1991) describe the NFCS and the selection of these women; they also tabulate summary statistics for self-reported information: body weight and the consumption of all foods, beverages, and water for 3 consecutive days. These are the most current data now available.

Table 1 presents the empirical summary statistics computed by Ershow et al. (1991): the arithmetic mean, the arithmetic standard deviation, and key percentiles for the amount of total water and tap water ingested daily by the 3 groups of women. Total water intake equals tap water intake (including, coffee, tea, and other beverages or foods made from or reconstituted with tap water) plus other water intake (including carbonated beverages, most alcoholic beverages, and intrinsic water in foods). In their original publication, Ershow et al. presented their results in two forms: (i) the water intake per day (denoted here by WI; measured in g/day and (ii) the water intake per day normalized to body weight (denoted here by WI/BW; measured in g/(kg•day)).

We agree with the US Environmental Protection Agency's selection of Ershow et al.'s summaries of the NFCS Survey as a good source of information on the amount of total water and tap water ingested by these 3 groups of women (US EPA, 1996). In this manuscript, we fit parametric (LogNormal) distributions to Ershow et al.'s summaries of the empirical data (reprinted in Table 1) so that risk assessors may use the fitted distributions efficiently in probabilistic exposure assessments.

<u>Methods and Results</u>

After completing an exploratory data analysis (Tukey, 1977), we fit LogNormal distributions to the data in Table 1 using these parameterizations (Evans et al, 1993):

$$\mathrm{WI} \sim \exp[\ \mathrm{Normal}[\ \mu_{\mathrm{WI}},\ \sigma_{\mathrm{WI}}\ ]\ ] \qquad \text{Eqn 1}$$

$$\frac{\mathrm{WI}}{\mathrm{BW}} \sim \exp[\ \mathrm{Normal}[\ \mu_{\frac{\mathrm{WI}}{\mathrm{BW}}},\ \sigma_{\frac{\mathrm{WI}}{\mathrm{BW}}}\ ]\ ] \qquad \text{Eqn 2}$$

Eqns 1 and 2 are equivalent, respectively, to $\ln[\ \mathrm{WI}\ ] \sim \mathrm{Normal}[\ \mu_{\mathrm{WI}},\ \sigma_{\mathrm{WI}}\ ]$ and $\ln[\ \frac{\mathrm{WI}}{\mathrm{BW}}\ ] \sim \mathrm{Normal}[\ \mu_{\frac{\mathrm{WI}}{\mathrm{BW}}},\ \sigma_{\frac{\mathrm{WI}}{\mathrm{BW}}}\ ]$ Here, exp[•] represents the exponential function, ln[•] represents the Napierian (or natural) logarithm function, and Normal[μ, σ] represents the Normal or Gaussian distribution with mean μ and standard deviation σ (with $\sigma > 0$).

First, we fit LogNormal distributions to the data using probability plots (D'Agostino & Stephens, 1986; Burmaster & Hull, 1996). Figure 1 shows LogNormal probability plots developed in Mathematica (Wickham-Jones, 1994) for the water ingested daily by each of the 3 groups of women. In the 6 panels, the small dots mark the key percentiles from Table 1 and the large dots mark the arithmetic means. The solid and dashed lines, respectively, indicate straight lines fit by ordinary least squares (OLS) regression to the data for total water intake and tap water intake. The first two columns of results in Table 2 give the intercept and the slope of the OLS lines in Figure 1 (respectively, the values for $\hat{\mu}$ and $\hat{\sigma}$ in Eqns 1 and 2). The next column in Table 2 gives the adjusted R^2 (aR^2) value for the regression. The high aR^2 values agree with the excellent visual fit of the model to the data as seen in Figure 1.

Second, we fit LogNormal distributions to the data using Maximum Likelihood Estimation (MLE) (Edwards, 1992). Since we do not have the individual measured data, we must use a cumulative distribution function (CDF) to develop the loglikelihood function for these "binned" data (e.g., Tanner, 1996, p. 15). The CDF for the univariate Normal distribution, denoted Φ, is:

$$\Phi[\, x \mid \mu, \sigma\,] \quad = \quad \frac{1}{2}\Big[\, 1 + \mathrm{erf}[\, \frac{x - \mu}{\sigma\sqrt{2}}\,]\,\Big] \qquad \text{Eqn 3}$$

where erf[•] denotes the error function (Abramowitz & Stegun, 1964). The likelihood function for the data in <u>one</u> of the bins between $\{WI_{lo}, p_{lo}\}$ and $\{WI_{hi}, p_{hi}\}$ is:

$$\Big[\, \Phi[\, \ln[WI_{hi}] \mid \mu, \sigma] - \Phi[\, \ln[WI_{lo}] \mid \mu, \sigma\,]\,\Big]\ .$$

Here, "lo" and "hi" refer to the bounds of a bin. The likelihood function for the WI measurements <u>in the bin</u> between $\{WI_{lo}, p_{lo}\}$ and $\{WI_{hi}, p_{hi}\}$ is:

$$\Big[\, \Phi[\, \ln[WI_{hi}] \mid \mu, \sigma] - \Phi[\, \ln[WI_{lo}] \mid \mu, \sigma\,]\,\Big] \text{ ^ } (\, N\, (p_{hi} - p_{lo})\,)\ .$$

The loglikelihood function for all N of the WI measurements, denoted $J[\, \mu, \sigma\,]$, is:

$$J[\, \mu, \sigma\,]$$

$$= \quad N \bullet \sum\, (\, p_{hi} - p_{lo}\,)\, \ln\Big[\, \Phi[\, \ln[WI_{hi}] \mid \mu, \sigma] - \Phi[\, \ln[WI_{lo}] \mid \mu, \sigma\,]\,\Big]\ . \qquad \text{Eqn 4}$$

for the sum over the bins. For the lowest bin, $\{WI_{lo}, p_{lo}\} = \{\varepsilon, \varepsilon\}$, and for the highest bin, $\{WI_{hi}, p_{hi}\} = \{1/\varepsilon, 1-\varepsilon\}$, with $\varepsilon = 10^{-5}$.

The 3 columns on the right side of Table 2 give the values for $\hat{\mu}$, $\hat{\sigma}$, and the maximum of the loglikelihood function found by direct maximization of Eqn 4 in Mathematica (Wolfram, 1991). The parameters $\hat{\mu}$ and $\hat{\sigma}$ concisely summarize the information in the fitted LogNormal distributions. For example, mode = exp[$\hat{\mu} - \hat{\sigma}^2$], median = geometric mean = exp[$\hat{\mu}$], the arithmetic mean = exp[$\hat{\mu} + \frac{1}{2}\hat{\sigma}^2$], and the 95th percentile = exp [$\hat{\mu} + z_{0.95}\,\hat{\sigma}$].

Third, based on theory and on previous results (Brainard & Burmaster, 1992; Burmaster & Crouch, 1996), we postulate that the variables WI (in g/day) and BW (in kg) are jointly distributed according to this bivariate LogNormal distribution:

$$\begin{bmatrix} WI \\ BW \end{bmatrix} \sim \exp[\,\mathrm{Normal}[\,\mu_{WI},\ \sigma_{WI},\ \mu_{BW},\ \sigma_{BW},\ \rho\ ]\,] \qquad \text{Eqn 5}$$

which is equivalent to having the natural logarithms of the two variables jointly distributed according to this bivariate Normal distribution (Freund, 1971, Evans et al, 1993):

$$\begin{bmatrix} \ln[\,WI\,] \\ \ln[\,BW\,] \end{bmatrix} \sim \mathrm{Normal}[\,\mu_{WI},\ \sigma_{WI},\ \mu_{BW},\ \sigma_{BW},\ \rho\ ] \qquad \text{Eqn 6}$$

In Eqns 5 and 6, ρ is the Pearson correlation coefficient (Keeping, 1995) in logarithmic space.

In this bivariate distribution, WI and BW are the two marginal LogNormal distributions. When Eqns 5 and 6 obtain, the random variable WI/BW is also a LogNormal random variable:

$$\frac{WI}{BW} \sim \exp[\,\mathrm{Normal}[\,\mu_{\frac{WI}{BW}},\ \sigma_{\frac{WI}{BW}}\ ] \qquad \text{Eqn 7}$$

where (Mood et al, 1974):

$$\mu_{\frac{WI}{BW}} = \mu_{WI} - \mu_{BW} \qquad \text{Eqn 8}$$

and

$$\sigma^2_{\frac{WI}{BW}} = \sigma^2_{WI} + \sigma^2_{BW} - 2\,\rho\ \sigma_{WI}\ \sigma_{BW} \qquad \text{Eqn 9}$$

We now interpret the results in Table 2 in terms of an underlying bivariate LogNormal distribution. First, we estimate μ_{BW} for each of the 3 groups of women using Eqn 8 and the results in Table 2. For each of the 3 groups of women, the estimated values of μ_{BW}

in Table 3 for total water ingestion and for tap water ingestion agree closely. The median weights for the control women are consistent with the results from previous research (Brainard & Burmaster, 1992; Burmaster & Crouch, 1996) and the results for the pregnant and lactating women are consistent with medical knowledge. Next, we estimate the pairs { σ_{BW}, ρ } for each of the 3 groups of women using Eqn 9 and the results in Table 2. For each of the 3 groups of women, the estimated values of ρ show that WI and BW have a weak positive correlation for each of the 3 groups. According to the criterion ($|\rho| < 0.6$) published by Smith et al. (1992), the correlations of WI and BW are so weak that they are numerically unimportant in computer simulations of exposure, i.e., the correlations can be safely ignored without damaging the overall quality and reliability of the results.

Discussion and Conclusions

First, we agree with the US EPA that Ershow et al.'s summary of the NFCS Survey (Table 1) is a reliable source of information on the amount of total water and tap water ingested by 3 groups of women in the United States. (The US EPA has long relied upon results from Ershow and Cantor's earlier analyses for the general population (US EPA, 1996; Ershow & Cantor, 1989; Roseberry & Burmaster, 1992)). Since the USDA did design the Survey to select a representative sample of women from the general population (including different geographical regions), and since the USDA did not design the Survey to select or to reject pregnant or lactating women, there are no known systematic biases in the selection of the population. However, since the consumption of bottled water has increased in the United States since the USDA completed the Survey, the results here likely overstate current consumption patterns for home tap water.

Second, in general, these data show that (i) lactating women ingest more water than do pregnant women and (ii) pregnant women ingest more water than the control women. See also the results in Ershow and Cantor (1989).

Third, LogNormal distributions fit each of the of data sets well. The distributions for total water intake have a better fit than the distributions for tap water intake.

Fourth, the results in Table 2 from the two statistical methods (probability plots and MLE) agree to within a few percent for each group of women.

Fifth, the results in Table 2 for the control group are consistent with the previous results for drinking water ingestion by adult women (Ershow & Cantor, 1989; Roseberry & Burmaster, 1992) and with previous results for body weight of women (Brainard & Burmaster, 1992; Burmaster & Crouch, 1997).

Sixth, we recommend the (marginal) distributions for WI as fit by MLE for the *variability* in a population for use in short-term human health risk assessments and pharmacokinetic models. For the control, pregnant, and lactating groups, respectively, the standard "default" of 2 l/day for water ingestion falls at the 88th, 86th, and 86th percentiles of the fitted distributions for tap water intake.

Seventh, we note that the results here are consistent with having the variables WI and BW jointly distributed according to a bivariate LogNormal distribution with a small positive Pearson correlation. In this larger framework, the derived variable WI/BW also follows a LogNormal distribution for each group. Little or no precision is gained by normalizing water intake by body weight and serious mistakes can arise by normalizing water intake by body weight in multi-pathway exposure assessments (see discussion in Ferson, 1996; page 562 concerning "instantiations").

Finally, a caveat. All the results in Ershow et al (1991) rely on information self-reported by the Survey participants for 3 consecutive days. Since most risk assessors are more interested in long-term average exposures than in 3-day exposures, the LogNormal distributions fitted here for the variability in short-term data have expected values (i.e., arithmetic mean = $\exp[\,\hat{\mu} + \frac{1}{2}\,\hat{\sigma}^2\,]$) that closely approximate the long-term <u>average</u> exposure. Overall, the distributions applicable for long-term average exposures will have expected values similar to the expected values for the distributions for short-term exposures, but with shorter tails (plural), i.e., smaller variance.

<u>Acknowledgments</u>

Kara B. Altshuler and Abby G. Ershow provided many helpful comments during this research. Two anonymous reviewers made excellent suggestions for improving the manuscript. Alceon Corporation funded this research

<u>Trademarks</u>

Mathematica ® is a registered trademark of Wolfram Research, Inc: http://www.wri.com

Alceon ® is a registered trademark of Alceon Corporation: http://www.Alceon.com

References

Abramowitz & Stegun, 1964
Abramowitz, M. and I.A. Stegun, Eds, 1964, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, Applied Mathematics Series Number 55, Issued June 1964, Tenth Printing with corrections in December 1972, US Government Printing Office, Washington, DC

Brainard & Burmaster, 1992
Brainard, J. and D.E. Burmaster, 1992, Bivariate Distributions for Height and Weight of Men and Women in the United States, Risk Analysis, 1992, Volume 12, Number 2, pp 267 - 275

Burmaster & Crouch, 1997
Burmaster, D.E. and E.A.C. Crouch, 1997, Lognormal Distributions for Body Weight as a Function of Age for Males and Females in the United States, 1976 - 1980, Risk Analysis, in press

Burmaster & Hull, 1997
Burmaster, D.E. and D.A. Hull, 1997, Using LogNormal Distributions and LogNormal Probability Plots in Probabilistic Risk Assessments, Human and Ecological Risk Assessment, Volume 3, Number 2, pp 235 - 255

D'Agostino & Stephens, 1986
D'Agostino, R.B. and M.A. Stephens, 1986, Goodness-of-Fit Techniques, Marcel Dekker, New York, NY

Edwards, 1992
Edwards, A.W.F., 1992, Likelihood, Expanded Edition, John Hopkins University Press, Baltimore, MD

Ershow et al, 1991
Ershow, A.G., L.M. Brown, and K.P. Cantor, 1993, Intake of Tap Water and Total Water by Pregnant and Lactating Women, American Journal of Public Health, Volume 81, Number 3, pp 328 - 334

Ershow & Cantor, 1989
Ershow, A.G. and K.P. Cantor, 1989, Total Water and Tap Water Intake in the United States: Population-Based Estimates of Quantities and Sources, Federation of American Societies for Experimental Biology, Bethesda, MD

Evans et al, 1993
Evans, M., N. Hastings, and B. Peacock, 1993, Statistical Distributions, Second Edition, John Wiley & Sons, New York, NY

Ferson, 1996
Ferson, S., 1996, Automated Quality Assurance Checks on Model Structure in Ecological Risk Assessments, Human and Ecological Risk Assessment, Volume 2, Number 3, pp 558 - 569

Freund, 1971
Freund, J.E., 1971, Mathematical Statistics, Second Edition, Prentice-Hall, Englewood Cliffs, NJ

Keeping, 1995
Keeping, E.S., 1995, Introduction to Statistical Inference, Dover, New York, NY

Mood et al, 1974
Mood, A.M., F.A. Graybill, and D.C. Boes, 1974, Introduction to the Theory of Statistics, Third Edition, McGraw Hill, New York, NY

Roseberry & Burmaster, 1992
Roseberry, A.M., and D.E. Burmaster, 1992, Lognormal Distributions for Water Intake by Children and Adults, Risk Analysis, Volume 12, Number 1, pp 99 - 104

Smith et al, 1992
Smith, A.E., P.B. Ryan, and J.S. Evans, 1992, The Effect of Neglecting Correlations When Propagating Uncertainty and Estimating Population Distribution of Risk, Risk Analysis, Volume 12, Number 4, pp 467 - 474, December 1992

Tanner, 1996
Tanner, M.A., 1996, Tools for Statistical Inference, Third Edition, Springer-Verlag, New York, NY

Tukey, 1977
Tukey, J.W., 1977, Exploratory Data Analysis, Addison-Wesley, Reading, MA

USDA, 1980
US Department of Agriculture, 1980, Nationwide Food Consumption Survey 1977 - 1978 Individual Data, Report I-2, Hyattsville, MD

US EPA, 1996
US Environmental Protection Agency, 1996, Exposure Factors Handbook, Science Advisory Board Review Draft, Exposure Assessment Group, EPA/600/P-96/002B a, b, and c, Washington, DC, August 1996

Wickham-Jones, 1994
Wickham-Jones, T., 1994, Mathematica Graphics, Techniques & Applications, Springer-Verlag, Telos, Santa Clara, CA

Wolfram, 1991
Wolfram, S., 1991, Mathematica®, A System for Doing Mathematics by Computer, Second Edition, Addison- Wesley, Redwood City, CA

deb 30 Mar 97
proofed

Table 1
Data for
Water Intake of Women, 15 - 49 Years Old

Variable	Group	Source	N	Units for Ingestion	Arithmetic Mean	Arithmetic StdDev	5th Percentile	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile	95th Percentile
•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••
WI	Control	total	6201	g/day	1940	686	995	1172	1467	1835	2305	2831	3186
WI	Control	tap	6201	g/day	1157	635	310	453	709	1065	1503	1983	2310
WI	Pregnant	total	188	g/day	2076	743	1085	1236	1553	1928	2444	3028	3475
WI	Pregnant	tap	188	g/day	1189	699	274	419	713	1063	1501	2191	2424
WI	Lactating	total	77	g/day	2242	658	1185	1434	1833	2164	2658	3169	3353
WI	Lactating	tap	77	g/day	1310	591	430	612	855	1330	1693	1945	2191
WI/BW	Control	total	6201	g/(kg•day)	32.3	12.3	15.8	18.5	23.8	30.5	38.7	48.4	55.4
WI/BW	Control	tap	6201	g/(kg•day)	19.1	10.8	5.2	7.5	11.7	17.3	24.4	33.1	39.1
WI/BW	Pregnant	total	188	g/(kg•day)	32.1	11.8	16.4	17.8	22.8	30.5	40.4	48.9	53.4
WI/BW	Pregnant	tap	188	g/(kg•day)	18.3	10.4	4.9	5.9	10.7	16.4	23.8	34.5	39.6
WI/BW	Lactating	total	77	g/(kg•day)	37.0	11.6	19.6	21.8	28.4	35.1	45.0	53.7	59.2
WI/BW	Lactating	tap	77	g/(kg•day)	21.4	9.8	7.4	9.8	14.8	20.5	26.8	35.1	37.4

Source: Tables 2 and 3 (Ershow et al, 1991)

deb 30 March 97
proofed

Table 2
Best-Fit LogNormal Distributions for
Water Intake of Women, 15 - 49 Years Old

Variable	Group	Source	Units for Ingestion	ProbPlot muhat	ProbPlot sigmahat	ProbPlot aR2	MLE muhat	MLE sigmahat	MLE maxJ
•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••
WI	Control	total	g/day	7.505	0.349	0.9984	7.510	0.347	-11,561.7
WI	Control	tap	g/day	6.863	0.594	0.9801	6.906	0.593	-11,730.5
WI	Pregnant	total	g/day	7.570	0.351	0.9997	7.570	0.349	-350.2
WI	Pregnant	tap	g/day	6.856	0.646	0.9754	6.904	0.640	-358.9
WI	Lactating	total	g/day	7.658	0.310	0.9838	7.675	0.307	-146.0
WI	Lactating	tap	g/day	7.018	0.481	0.9550	7.073	0.492	-150.5
WI/BW	Control	total	g/(kg•day)	3.402	0.377	0.9989	3.408	0.375	-11,556.6
WI/BW	Control	tap	g/(kg•day)	2.762	0.595	0.9850	2.798	0.591	-11,687.0
WI/BW	Pregnant	total	g/(kg•day)	3.398	0.377	0.9946	3.406	0.384	-353.3
WI/BW	Pregnant	tap	g/(kg•day)	2.703	0.650	0.9881	2.740	0.646	-355.4
WI/BW	Lactating	total	g/(kg•day)	3.548	0.342	0.9963	3.557	0.342	-144.2
WI/BW	Lactating	tap	g/(kg•day)	2.924	0.489	0.9764	2.963	0.488	-147.0

deb 30 March 97
proofed

Table 3
Best-Fit Bivariate LogNormal Distributions for
Water Intake and Body Weight of Women, 15 - 49 Years Old

Group	Source	MLE WI muhat	MLE WI sigmahat	MLE WI/BW muhat	MLE WI/BW sigmahat	Implied BW muhat	Paired BW sigmahat	Paired WI, BW rhohat	AMean BW (kg)
•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••	•••••
Control	total	7.510	0.347	3.408	0.375	4.102	0.228	0.201	62.07
Control	tap	6.906	0.593	2.798	0.591	4.108			
Pregnant	total	7.570	0.349	3.406	0.384	4.165	0.217	0.142	65.90
Pregnant	tap	6.904	0.640	2.740	0.646	4.164			
Lactating	total	7.675	0.307	3.557	0.342	4.117	0.259	0.278	63.48
Lactating	tap	7.073	0.492	2.963	0.488	4.110			

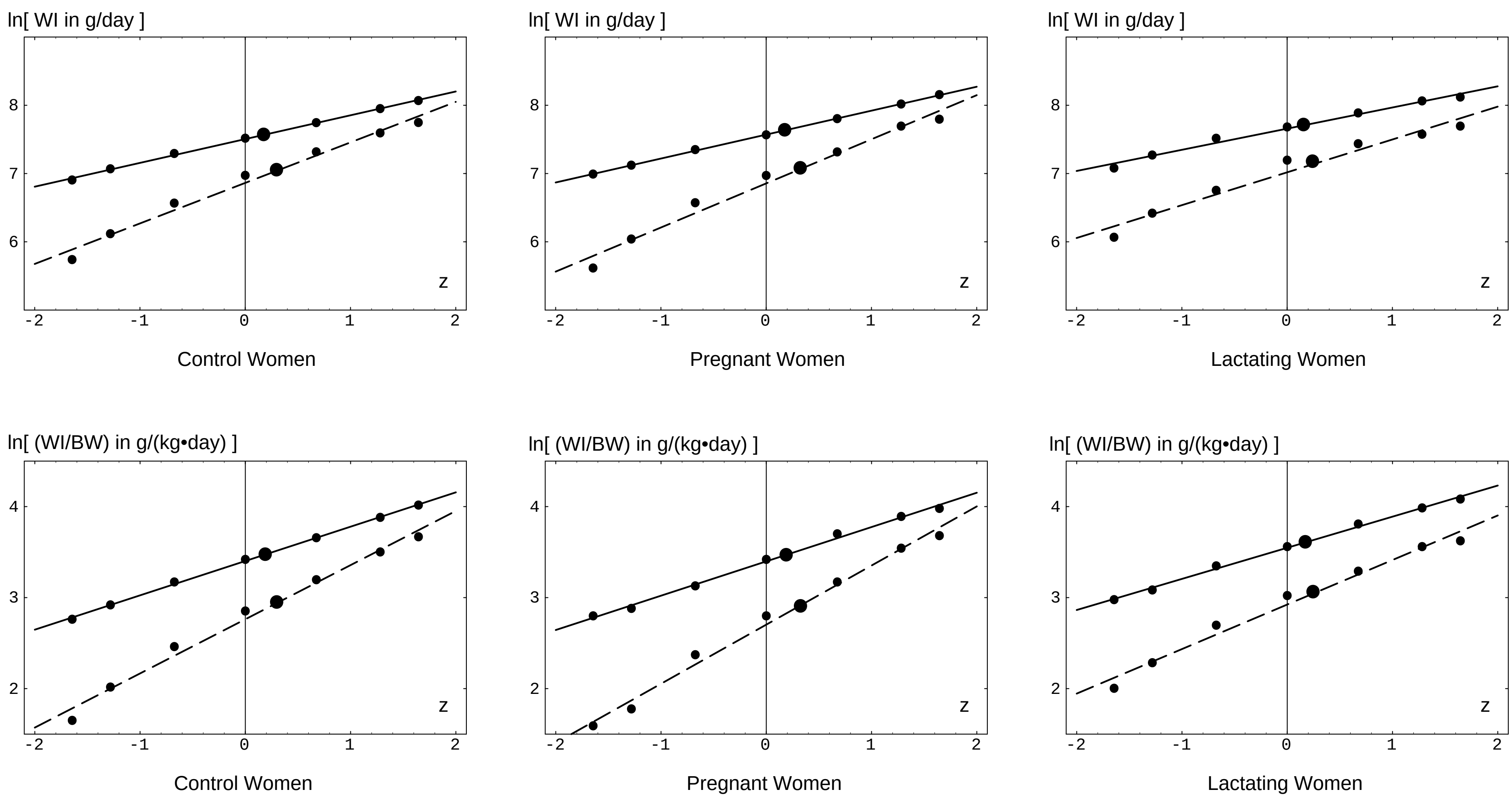

Figure 1
LogNormal Probability Plots for
Water Ingested by Women;
solid lines indicate total water intake and
dashed lines indicate tap water intake