Inference about Cryptosporidium

December 6, 1999

By **Robert L. Wolpert**

Institute of Statistics and Decision Sciences

and

Nicholas School of the Environment,

Duke University

Outline

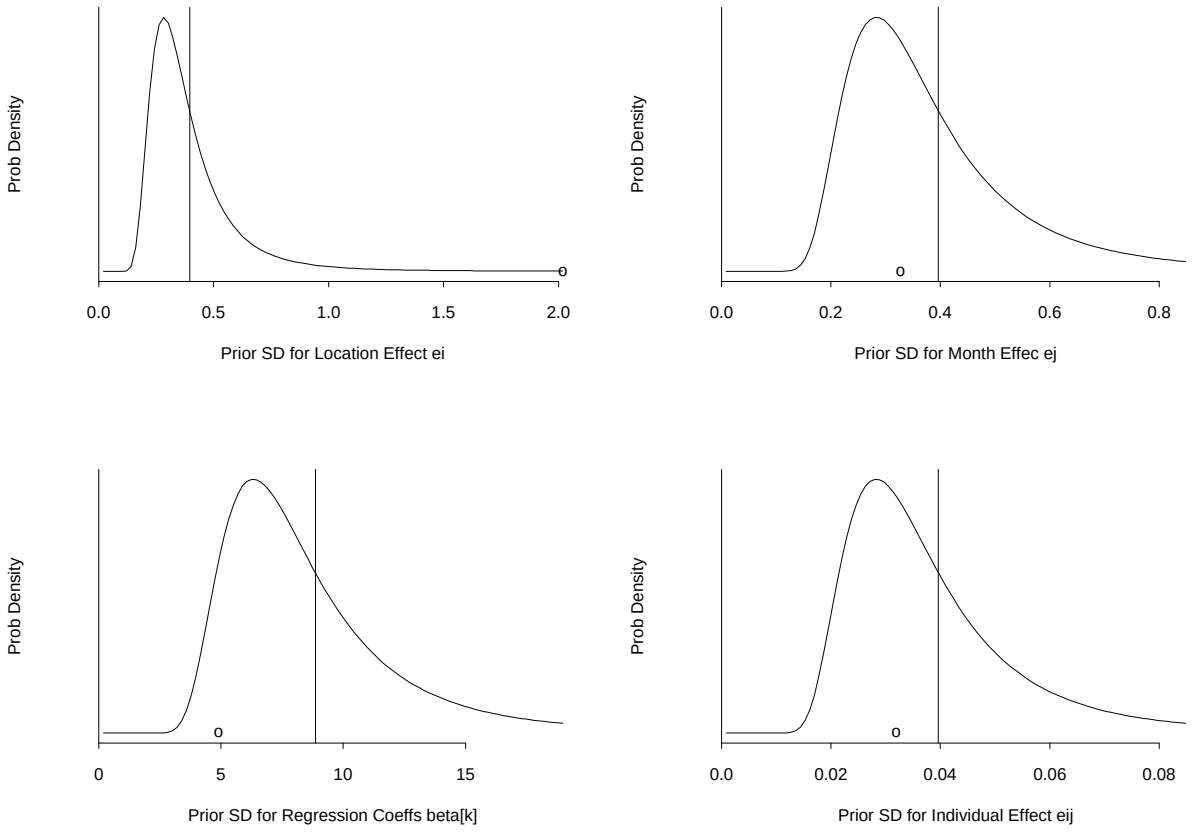
1. True Zeros & False Positives
2. Results
3. Conclusions and Future Work

1. True Zeros & False Positives

Counts:	Y_{ij}	$\sim$	$\mathsf{Po}(\lambda_{ij} = \eta_{\mathsf{FP}} + v_{ij} r_{ij} c_{ij})$
Volume:	v_{ij}	$=$	assumed known
Recovery:	r_{ij}	$\sim$	$\mathsf{Be}(\alpha_1 = 1.44, \alpha_2 = 11.2)$
False Positive:	η_{FP}	$=$	$1\% \times \dfrac{753}{3901} \approx 0.002$ (assumed known)
NonZero:	Z_i	$=$	$\begin{cases} 0 & \text{w/probability } p_{\mathsf{TZ}} \\ 1 & \text{w/probability } 1 - p_{\mathsf{TZ}} \end{cases}$
True Zeros:	p_{TZ}	$=$	0.10 (assumed known)
Concentration:	c_{ij}	$\equiv$	$Z_i \exp(\epsilon_i + \epsilon_j + \epsilon_{ij} + \sum_k X_{ijk} \beta_k)$

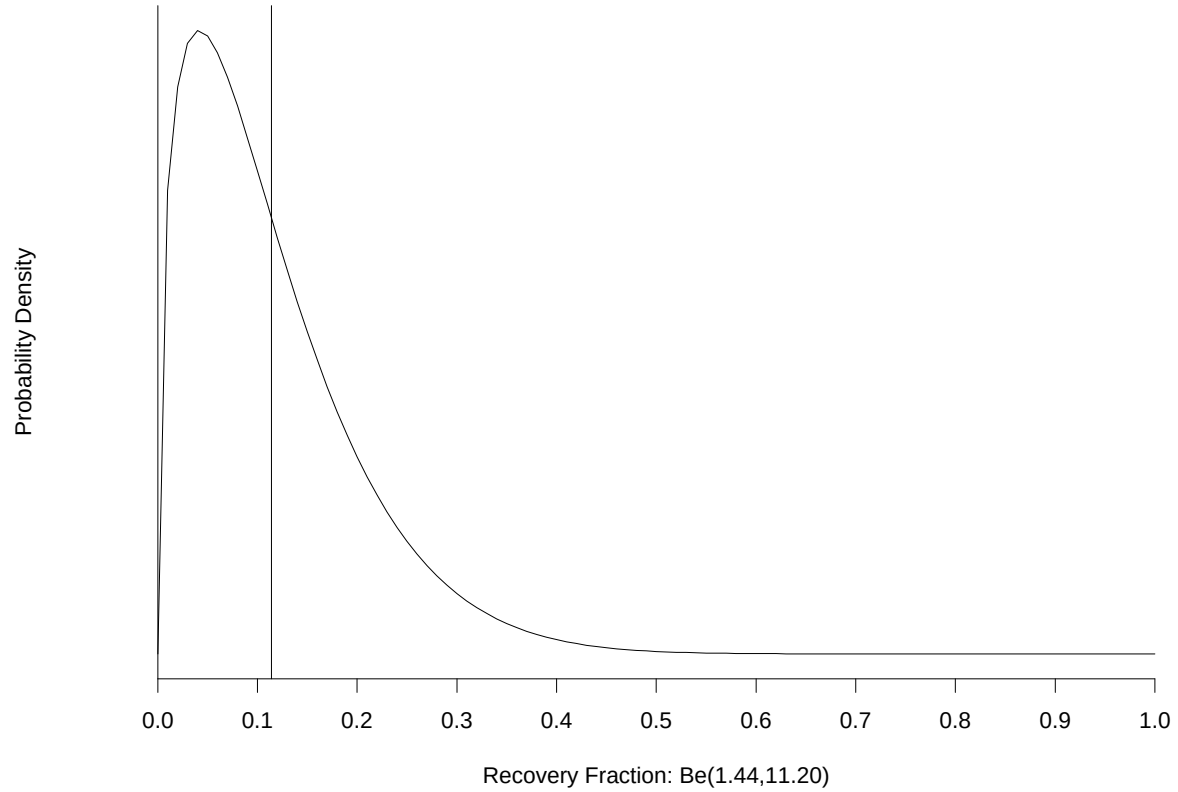
Location effect:	ϵ_i	$\sim$	$\mathsf{No}(0, \theta_1),\ i \in I$
Month effect:	ϵ_j	$\sim$	$\mathsf{No}(0, \theta_2),\ j \in J$
Residual:	ϵ_{ij}	$\sim$	$\mathsf{No}(0, \theta_3),\ ij \in I \times J$
Coefficients:	β_k	$\sim$	$\mathsf{No}(\mu_k, \theta_4),\ k \in \{1, 2, 3, 4, 5, 6, 7\}$
Intercept:	X_{ij1}	$=$	1
Turbidity:	X_{ij2}	$=$	Log Turbidity
MSW Class:	X_{ijk}	$=$	1 for RL, UF, GI, MX, UK, $k = 3:7$
			($\Longrightarrow$ β_k is difference from FS)
Precisions:	θ_k	$\sim$	$\mathsf{Ga}(2.0, \tau_k),\ k \in \{1, 2, 3, 4\}$
			($\Longrightarrow$ $\epsilon_i, \epsilon_j, \epsilon_{ij}, \beta_k \sim t_4$)

Prior Densities for SD's from Precisions:



Implicit Priors for $\tau_k^{-1/2}$, if $\tau_k \sim \mathsf{Ga}(\alpha_k, \tau_k)$

Recovery Fraction:

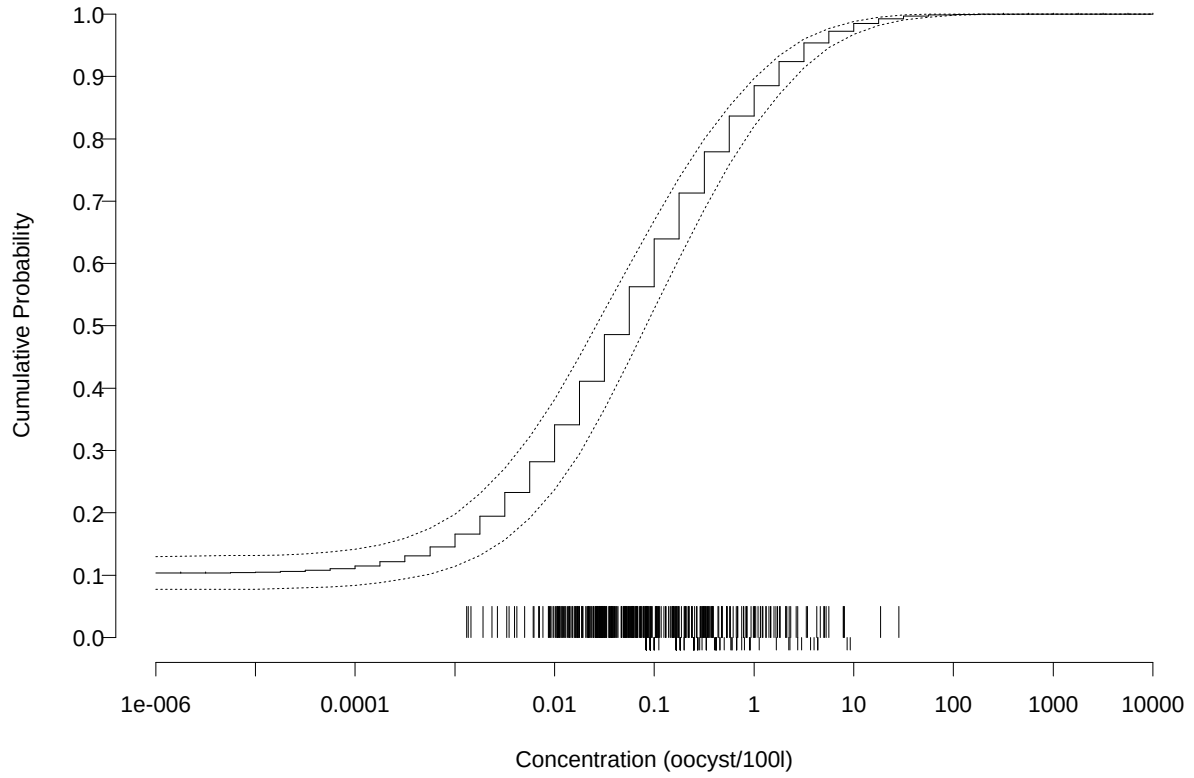


Taken from Spiking Studies (not critical)

False Positives, Protozoa-Free Probabilities:

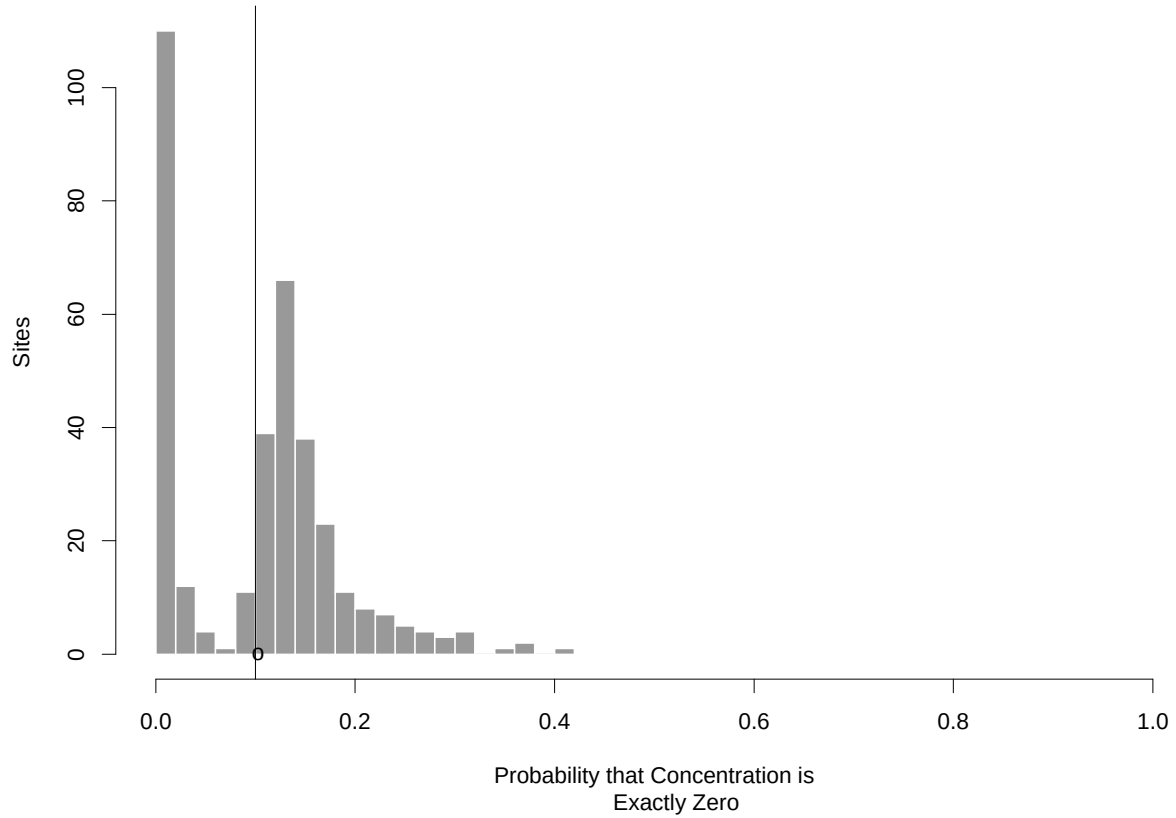
- 753 Oocysts in 3901 records; one percent false positives would give $\eta_{\mathsf{FP}} \approx (0.01 \times 753)/3901 = 0.00193$
- 223 of 350 Sites (63.7%) have no reported oocysts in 12 months $\Longrightarrow p_{\mathsf{TZ}} \ll 0.64$
- 51 of 350 Sites (14.6%) have one reported oocyst in 12 months
- Initial estimate: $\eta_{\mathsf{FP}} = 0.002$, $p_{\mathsf{TZ}} = 0.10$

2. Results (FP=0.01, TZ=0.10):



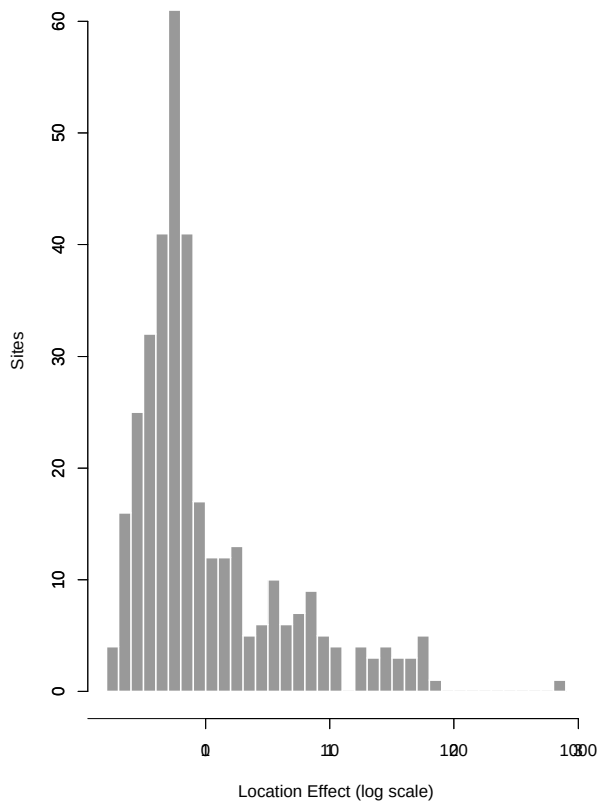
CDF of Posterior Site Means $\tilde{c}_i = \mathsf{E}[\frac{1}{12}\sum_j c_{ij}]$

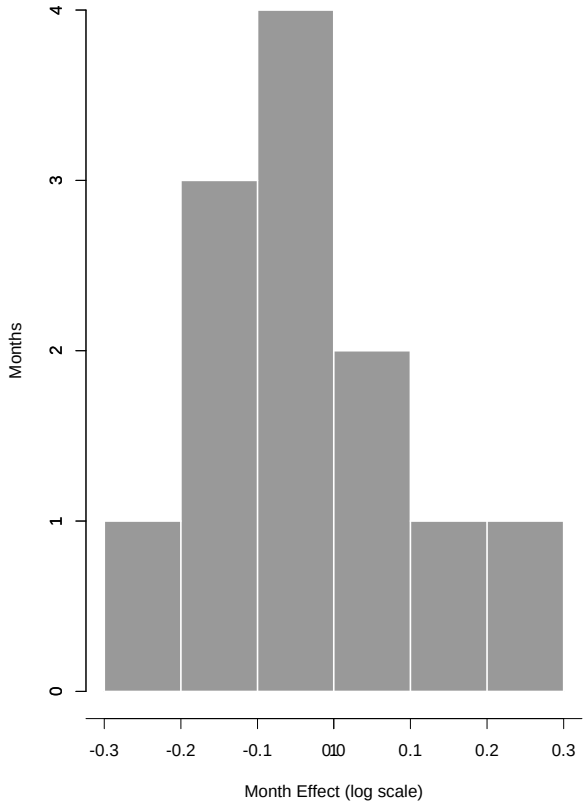
Probability of Exact Zero:



Probability $\mathsf{P}[Z_i = 0 \mid \text{Data}]$ (Prior mean 0.10)

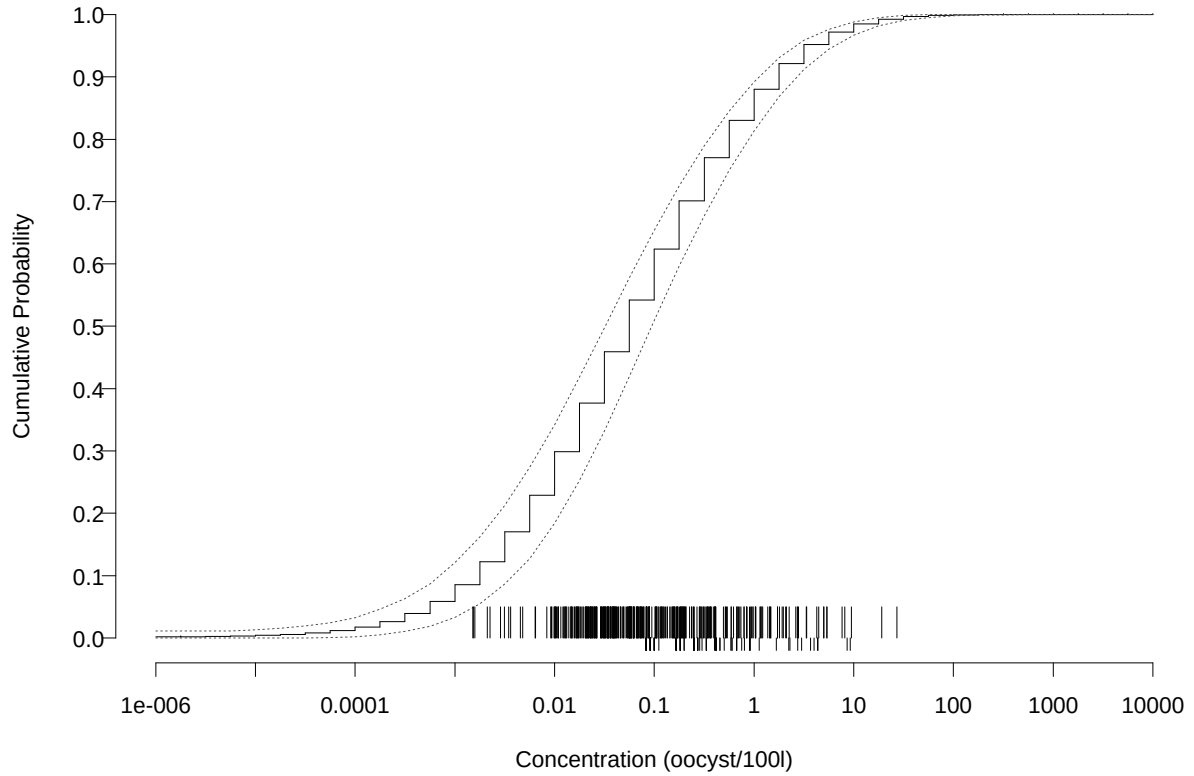
Individual and Month Effects: (FP=0.01, TZ=0.10):





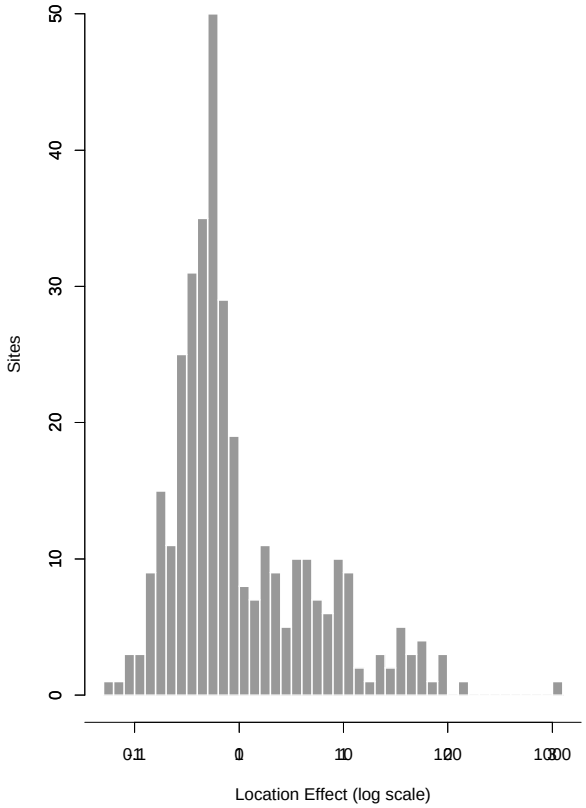
Location Effects $\mathsf{E}[\epsilon_i \mid \text{Data}]$

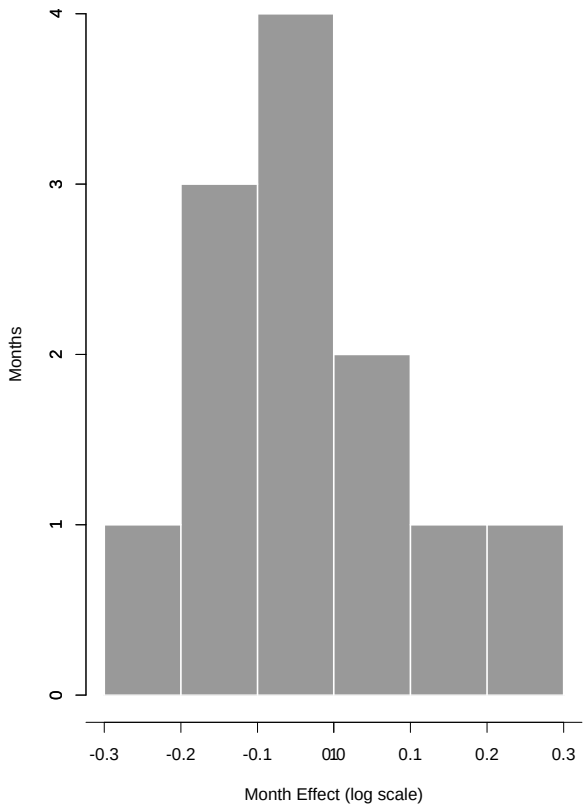
Results (FP=0, TZ=0):



CDF of Posterior Site Means $\tilde{c}_i = \mathsf{E}[\frac{1}{12}\sum_j c_{ij}]$

Individual and Month Effects: (FP=0, TZ=0):





Location Effects $\mathsf{E}[\epsilon_i \mid \text{Data}]$, Month Effects $\mathsf{E}[\epsilon_j \mid \text{Data}]$

Regression Coefficients: (FP=0.01, TZ=0.10):

				90% Interval		Factor
β_1	Intercept	-1.2985	$\in$	$[-1.4888,$	$-1.2182]$	0.0500
β_2	Turbidity	0.2204	$\in$	$[0.1959,$	$0.2476]$	1.7000
β_3	RL - FS	-0.6723	$\in$	$[-0.8738,$	$-0.5042]$	0.2100
β_4	UF - FS	-1.5348	$\in$	$[-2.1102,$	$-0.8768]$	0.0290
β_5	GI - FS	-2.2918	$\in$	$[-5.9715,$	$-0.5255]$	0.0051
β_6	MX - FS	-1.9647	$\in$	$[-2.2431,$	$-1.7024]$	0.0110
β_7	UK - FS	-0.3684	$\in$	$[-0.6597,$	$-0.1105]$	0.4300

Regression Coefficients: (FP=0, TZ=0):

				90% Interval		Factor
β_1	Intercept	-1.2712	$\in$	$[-1.4375,$	$-1.1513]$	0.2805
β_2	Turbidity	0.2313	$\in$	$[0.1922,$	$0.2581]$	1.2600
β_3	RL - FS	-0.6236	$\in$	$[-0.7344,$	$-0.4725]$	0.5360
β_4	UF - FS	-1.4323	$\in$	$[-2.0911,$	$-0.7539]$	0.2388
β_5	GI - FS	-2.1376	$\in$	$[-4.4645,$	$-0.3925]$	0.1179
β_6	MX - FS	-1.8944	$\in$	$[-2.1432,$	$-1.6369]$	0.1504
β_7	UK - FS	-0.4065	$\in$	$[-0.6897,$	$-0.1067]$	0.6660

3. Conclusions

- Possibility of Protazoa-Free sources: small effect for $C > 10^{-2}$
- Possibility of False Positives: small effect for $C > 10^{-2}$
- Both have larger effects at low concentrations $C \ll 10^{-2}$
- Model comparison will help inform low-concentration estimates
- More data will also help inform low-concentration estimates
- PE data may inform choice of False Positive rates
- Month effect (if any) is small
- More exploration needed

Current Work

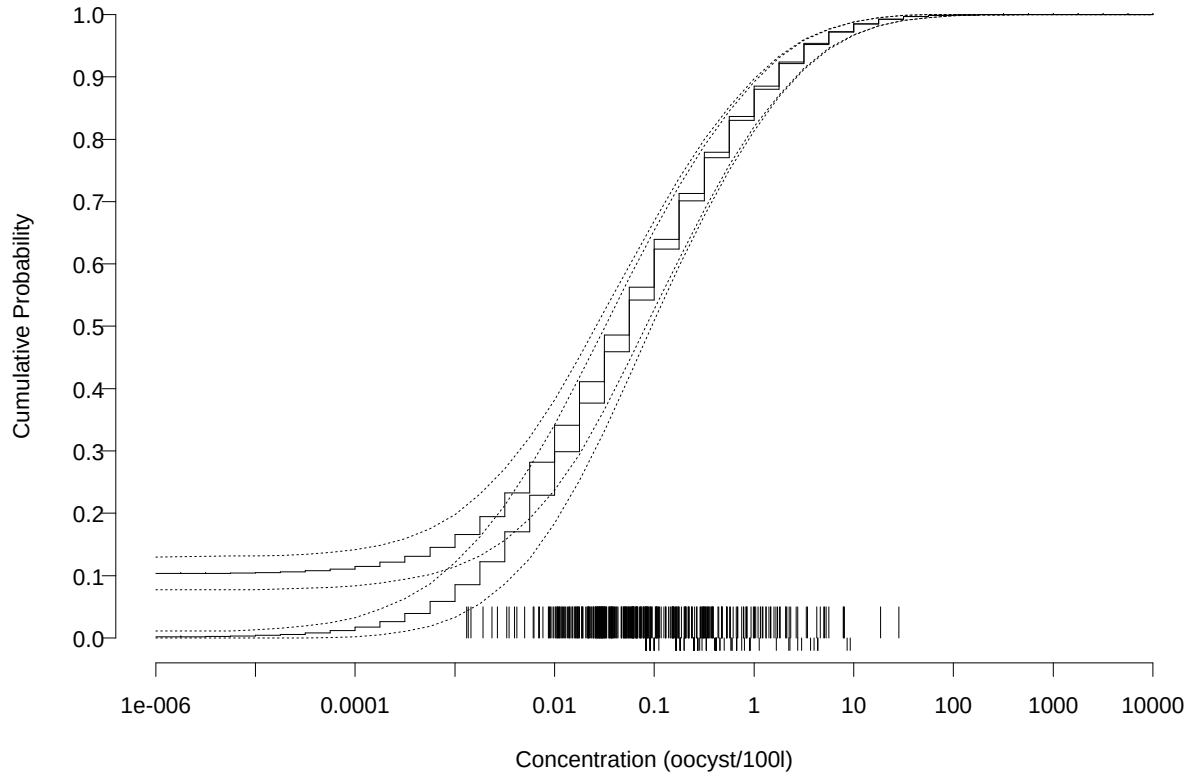
Recently completed or currently under way are:

- Exploring Turbidity
- Possibility of Protazoa-free Sources
- Possibility of False Positive Counts

Future Work

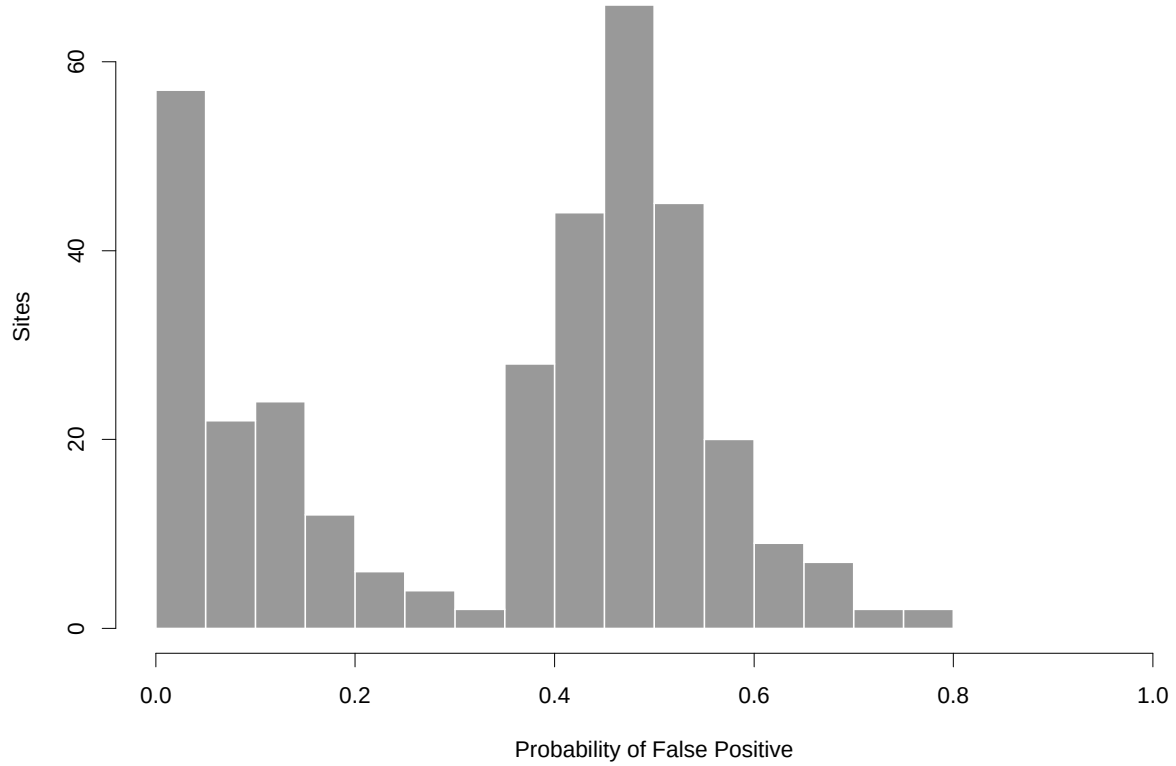
- Seasonality (replacing Months)
- Explore Spatial Effects (HUC?)
- Prior Elicitation
- Sensitivity Analysis & Model comparison (η_{FP}, p_{TZ})
- Giardia

Results (Both FP=0, FP=0.01):



CDF of Posterior Site Means $\tilde{c}_i = \mathsf{E}[\frac{1}{12}\sum_j c_{ij}]$

Probability of False Positives: (FP=0.01, TZ=0.10):



Posterior Probability of False Positives

2. Clumping

Let Y_{ij} be the number of microbes in a "pellet", and let X_{ij} be the number in a sample selected for analysis. If $Y_{ij} \sim \mathsf{Po}(\lambda)$ has a Poisson distribution with mean λ, and if the sample is some fraction p of the pellet, and if the pellet is well-mixed, then the number X_{ij} will also have a Poisson distribution but with mean $p\,\lambda$.

If the pellet is *not* well-mixed, but instead exhibits "clumping", then X_{ij} will still have *mean* $\mathsf{E}[X_{ij}] = p\,\lambda$ but will be have a distribution more widely spread out than that of a Poisson random variable— so it will be more likely to have quite small values (like zero) or large ones (well above $p\,\lambda$). One way to model this situation is to accord X_{ij} a "beta binomial" distribution, conditional on Y_{ij}, so that

$$\mathsf{P}[X_{ij} = k] = \int_0^1 \frac{\Gamma(\alpha+\beta)p^{\alpha-1}(1-p)^{\beta-1}}{\Gamma(\alpha)\,\Gamma(\beta)} \frac{(p\lambda)^k}{k!} e^{-p\,\lambda}\,dp$$