beginning to talk about statistics

statistical inference is based on a statistical model

statistical model relates observations to parametric scientific model (i.e., a model that is fully specified up to the value of parameters)

goal is to estimate parameters (fitting) and, more importantly, to characterize uncertainty in parameters (interval estimation) or test hypothesis about parameters (hypothesis testing) – it will turnout that interval estimation and hypothesis testing are dual

statistical model distinguished from deterministic model by some kind of randomness or stochasticity in observations

sources of randomness include: measurement error, 'natural variability' (unmodeled variability), sampling variability

observations are random variables, each characterized by a probability distribution which depends on the parameters of the underlying scientific model

example: salps

salps are disgusting jelly-like creatures that live in the ocean

let $\mu(t)$ be the mean density (# per unit volume) of salps in an area at time *t*

population dynamics follows the model:

$$\frac{d\mu(t)}{dt}\frac{1}{\mu(t)} = \beta\, x(t)$$

where *x*(*t*) is known temperature anomaly* at time *t*

$\rightarrow \boldsymbol{\mu(t) = \mu(0)\, exp(\beta \int_0^t x(u)du}$

2 parameters are $\mu(0)$ and β → observe $\mu(t)$ at just 2 times, you can <u>solve</u> for these parameters

in practice, don't observe $\mu(t)$ but random variable *Y*(*t*) = #salps in water sample of unit volume at time *t*

randomness here due to sampling variability – take a different sample, get a different count

in this case, if you have observations $Y(t_1), Y(t_2), \ldots, Y(t_n)$ with n > 2, then you cannot find parameter values so that the observations are fit exactly

in fact, if all you want are <u>point estimates</u> of the parameters, you can often find good ones in an *ad hoc* way (vary parameters, judge fit be eye) or slightly more formally by <u>least squares</u> (i.e., minimizing sum of squared differences between model fit and observations); but for inference generally you will need a statistical model

we will come back to the salp example until you are heartily sick of it

in a statistical model, the observations (which are random variables) have distributions whose parameters (e.g., means) depend on the parameters of the underlying scientific model; you need to think about this dependence but you also need to think about what kind of distribution the observations have

<u>some random variables and their distributions</u>

<u>discrete random variables</u>

discrete random variable takes only a countable number of values (e.g., number of salps in a water sample of unit value)

discrete random variable *Y* has a <u>probability mass function</u> (pmf):

$$p(y) = prob(Y = y)$$

convention is that random variable denoted by upper case letter and its <u>realized value</u> by the same letter in lower case

pmf satisfies $p(y) \geq 0$ and $\sum_y p(y) = 1$

Poisson:

$$p(y) = \frac{\theta^y \, exp(-\theta)}{y!} \qquad y = 1, 2, \ldots$$

one parameter: positive real number θ

if 'events' (like salps) occur randomly in time (or space) with a mean rate of θ, the number of events in an interval of length (or volume) *v* has a Poisson distribution with parameter $\theta \, v$

for Poisson random variable, mean = variance = θ; common statistical problem is modeling dependence of Poisson mean on explanatory variables

Binomial:

$$p(y) = \binom{n}{y} \, p^y \, (1 - p)^{n-y} \qquad y = 0, 1, \ldots, n$$

two parameters: positive integer *n* (typically <u>known</u>), probability 0 < *p*< 1

Y* is the number of 'successes' in *n* independent trials each with success probability *p

for binomial random variable: mean = *n p*, variance = *n p* (1 – *p*); common statistical problem is modeling dependence of binomial probability on explanatory variables – dose-response models relate probability of response (like death) to the dose of a toxic substance – here, *n* individual organisms are subjected to doses $x_1, x_2, \ldots, x_k$ and the number $Y(x_j)$ responding is assumed to have a binomial distribution with *n* trials and success probability $p(x_j)$ – interest in inference under a parametric model (usually, linear logistic regression model)- we'll see this later

it is often the case that count data are <u>over-dispersed</u> in relation to the Poisson distribution (the sample variance is quite a lot larger than the sample mean) –this can happen if events tend to cluster

(in general, the sample mean for observed values $y_1, y_2, \ldots, y_n$ is:

$$\bar{y} = \sum_{j=1}^{n} y_j \,/n$$

and the sample variance is:

$$s^2 = \sum_{j=1}^{n} \left(y_j - \bar{y}\right)^2 /(n-1))$$

also success count data are often over-dispersed in relation to the binomial distribution – this can happen if the success probability varies randomly

the <u>negative binomial</u> is a 2-parameter distribution for count data that accommodates extra-Poisson variability – we'll get to an example (not involving salps) later

the <u>beta-binomial</u> plays the same role for extra-binomial variability

in general, unaccounted over-dispersion doesn't screw up point estimation of the mean but gives an over-optimistic picture of uncertainty about the mean

I will talk about detection of over-dispersion and a nice *ad hoc* way of dealing with it later

<u>continuous random variables</u>

continuous random variable takes values on all or part of the real line (uncountable)

(cumulative) distribution function (c)df:

$\boldsymbol{F(y) = prob(Y \leq y)}$

$\boldsymbol{F(-\infty) = 0}$ **(or** $\boldsymbol{F(0) = 0}$ **for positive random variable)**
$\boldsymbol{F(\infty) = 1}$
$\boldsymbol{F}$ **is non-decreasing**

probability density function (pdf):

$\boldsymbol{f(y) = dF(y)/dy \neq prob(Y = y)}$ **indeed** $\boldsymbol{prob(Y = y) = 0}$

$\boldsymbol{f(y) \geq 0}$ **NB *f*(y) can be > 1**
$\boldsymbol{\int_{-\infty}^{\infty} f(y)dy = 1}$

also:

$\boldsymbol{F(y) = \int_{-\infty}^{y} f(y)dy}$

get probabilities from integrals of pdf corresponding to differences in df:

$$prob(a \leq Y \leq b \leq) = \int_a^b f(y)dy = F(b) - F(a)$$

the '*p*-quantile' of a random variable is the value *y* of the random variable satisfying $F(y) = p$ (that is, the probability of being less than or equal to *y* is *p*; the 0.5-quantile is called the median; sometimes use 'upper *p*-quantile' to mean the value *y* such that the probability of being greater than or equal to *y* is *p*; the upper *p*-quantile is equal to the (lower) $(1 - p)$-quantile

<u>uniform distribution</u>

a random variable is uniformly distributed between *a* and *b* (denote by *U*(*a*, b)) if its values are restricted to the interval (*a*, *b*) and all values in the interval are equally likely

$f(y) = \frac{1}{b-a}$ $\quad a \leq y \leq b$ **and 0 otherwise**

$$F(y) = \int_a^y f(y)dy = \frac{y-a}{b-a}$$

2 parameters, *a* and *b*; mean is $(a + b)/2$, variance is $(b - a)^2/12$

application: sightings or fossil finds of a species commonly assumed to be uniformly distributed between a lower bound (e.g., origination time) and upper bound corresponding to extinction time; interest usually centers on estimating extinction time or, for modern species, testing that extinction has occurred; uniform model wouldn't be appropriate if rate of sightings varied systematically over time

special case is *U*(0, 1); has a role in simulation (which we will get to)

exponential distribution

$$f(y) = \theta \, exp(-\theta \, y) \quad y \geq 0$$

$$F(y) = 1 - exp(-\theta \, y)$$

1 parameter θ; mean is $1/\theta$, variance is $1/\theta^2$

application: commonly used as a model for duration (e.g., times between events, lifetimes); special case of gamma distribution which is a flexible 2-parameter distribution for positive random variables

Aside: sometimes you want to find the distribution of a function of random variable with known distribution; here is an example of how you do this; suppose that *X* has a *U*(0, 1) distribution and you want to find the distribution of $Y = -log \, X$

in this class, log means natural log

here are the steps:

$$F(y) = prob(Y \leq y)$$

$$= prob(-log \, X \leq y)$$

$$= prob(X \geq exp(-y))$$

$$= 1 - prob(X < exp(-y))$$

$$= 1 - exp(-y)$$ **from df of *U*(0, 1) distribution**

which is df of exponential random variable with $\theta = 1$

(this is more than an exercise: it's the basis of Fisher's method for combining significance levels)

<u>normal distribution</u>

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{1}{2\sigma^2}\,(y-\mu)^2\right) \quad -\infty \le y \le \infty$$

familiar 'bell-shaped curve'

$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$ **where Φ is the cdf of *N*(0, 1) random variable**

notation $N(\mu, \sigma^2)$ means 'normal with mean μ and variance σ^2)

so you can get $N(\mu, \sigma^2)$ from Φ which is tabulated and on-line

2 parameters: $\mu = E(Y), \sigma^2 = Var\ Y$ (not every distribution is parameterized by mean, variance (e.g., exponential)

normal distribution commonly used for measurement error, commonly misused for EVERYTHING (e.g., normal-based methods like ANOVA are used for everything)

Central Limit Theorem <u>very roughly</u> says that sums of large number of random variables tend to have a normal distribution – e.g., binomial distribution is approximately normal for large *n*; sample mean, too

<u>lognormal distribution</u>

***Y* is lognormal if log *Y* is normal**

parameterized by mean, variance μ, σ^2 of log *Y*

$E(Y) = exp(\mu + \frac{\sigma^2}{2})$ mean of <u>linear</u> function of random variables is the same function of the means of the random variables, <u>not</u> true of non-linear functions

there is an expression for $Var\ Y$ involving μ, σ^2

lognormal commonly used for positive random variables like size; if a random variable represents the product of random variables …

<u>fun with quantiles</u>

$Y \sim N(\mu, \sigma^2)$, what is upper 0.05-quantile?

$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$
$\Phi(1.65) = 0.95$
$\frac{y-\mu}{\sigma} = 1.65 \rightarrow y = \mu + 1.65\,\sigma$ is the upper 0.05-quantile

in statistics, commonly use χ^2 distribution (not as model); this is 1 parameter distribution; integer parameter is called 'degrees of freedom'; the χ^2 distribution with *df* degrees of freedom has mean *df* and variance 2 *df*; it is the distribution of the <u>sum</u> of the squares of *df* independent $N(0, 1)$ random variables; its cdf is tabulated in books and on-line

suppose you're interested in the upper 0.05-quantile of the χ^2_{df} distribution; by the Central Limit Theorem, for large *df*, this distribution is approximately $N(df, 2\ df)$ and we worked out above the upper 0.05 quantile of this distribution

for example, let's take *df* = 1 (where the normal approximation should not betoo good!); the upper 0.05-quantile based on the

normal approximation is $df + \sqrt{2\,df}\;1.65 = 1 + 1.65\sqrt{2} = 3.3$;
using on-linecalculator (danielsoper.com/statcalc3/calc.aspx?id=12)
the correct answer is 3.84

Homework: Repeat this for *df* = 2 and 10; comment