MDSC 643.01, Biostatistics I
Brant, Fall 2002

PROBABILITY DISTRIBUTIONS

Tabulating Binary Outcomes: The Binomial Distribution

Random sample of size n from some population (size N?)
Dichotomous (two possible values) outcome, S(uccess) or F(ailure)
X = #S's out of n = a binomial random variable
p = Probability{S on a single "draw"} = proportion of S's in population

example: Sampling from ...

	FH-yes	FH-no
GM-yes	1,000,000	1,000,000
GM-no	2,000,000	96,000,000

For n not too small, $Prob\{X = k\} \approx \binom{n}{k} p^k (1-p)^{n-k}$

- probability (mass) function of the Binomial distribution
- X = binomial random variable (approximate)

Probability histogram

- relative freqency histogram for hypothetical infinite population
- bar graph with bar heights $Prob\{X = k\}$

Note: if small population size relative to n use Hypergeometric distribution

Tail areas: $Prob\{X \geqslant k\}, Prob\{X \leqslant k\}, Prob\{X > k\}, Prob\{X < k\},$

- cumulative distribution function

Summarizing the Distribution of a Random Variables

Hypothetical repeated samples: repeat sampling "experiment" (fresh start each time), resulting in a "meta-sample" of observed X values. Consider "limiting value" as meta-sample size approaches infinity

Expected value (a.k.a. Mean Value) = $E(X)$
= sequential limit of sample mean X values
(also = mean in probability histogram)

Variance (of a Random Variable = $Var(X)$
= sequential limit of sample variance, (sd)2, in progressively larger
gives rise to Standard Deviation $SD(X) = \sqrt{Var(X)}$

When X is a binomial random variable:
$E(X) = n \times p$
$Var(X) = n \times p \times (1-p)$
$SD(X) = \sqrt{n \times p \times (1-p)}$

e.g.

Counting Rare Events: The Poisson Distribution

e.g. X = # events of certain type occuring "randomly" in time or space
e.g. # particle (sub-atomic) emissions from radioactive object
e.g. # cases of rare (sporadic) disease in a locale in given period

- another **discrete** distribution, i.e. range (of possible values) is a discontinuous set of numbers

Based on mathematical arguments, probability function for such random variables = $Prob(X = k) = \frac{e^{-\mu}\mu^k}{k!}$, where μ is the Expected Value of X.

e.g.

Mean value and variance: $E(X) = \mu \,;\, Var(X) = \mu$

Poisson approximation to Binomial

X binomial n large, p small (R.O.T.= n>=100, p<=.01)
then $Prob(X = k) = \frac{e^{-np}(np)^k}{k!}$
i.e. Poisson distribution with μ = np

e.g.

Continuous Distributions

If X takes on all possible values in a continuous range must use "calculus-based" ideas to define probability distribution

Probability density function (p.d.f.): A smooth curve taking values >= 0.
Areas under the curve give probabilities for corresponding events.

example: Exponential distribution - p.d.f. $f(x)=\lambda e^{-\lambda x}$

- describes distribution of X=waiting time to random event (e.g. failure of a computer chip) where λ is the event rate

e.g.

The Normal Distribution

So far have regarded normal curve as an approximation to histogram.

If X is a continuous random variable with normal shaped p.d.f., $f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ then X is called a normal random variable.

$$E\ X=\mu\,,Var\ X=\sigma^2\,,SD(X)=\sigma$$

Calculating normal tail areas – conversion to $z=\frac{x-\mu}{\sigma}$

Normal approximation to Binomial distribution

If X Binomial, n large, p not too small. (R.O.T. np(1-p) >=5)

Prob (a <= X <= b) approx area under normal curve with $\mu=np$, $\sigma=\sqrt{n\,p(1-p)}$

e.g.