

STANDARD DISTRIBUTIONS - Examples/Extensions

GENETIC LINKAGE and MAPPING

- **Linkage Phase** - chromatid associations of alleles of linked loci
- same chromosome =coupled, different =repulsion
- **Genetic Recombination** - define R.F. (gametes or phenotypes); homologous case - greater the distance, greater chance of recombining. High **interference** = problem for multiple locus models. R.F. between loci **not additive**. Need **Mapping Function**
- **Haldane's Mapping Function**

Assume crossovers occur randomly along chromosome length and average number = λ , model as **Poisson**

$$P\{\text{No crossover}\} = e^{-\lambda} \quad \text{and} \quad P\{\text{Crossover}\} = 1 - e^{-\lambda}$$

1

Examples - continued

- $P\{\text{recombinant}\} = 0.5 \times P\{\text{Crossover}\}$ (each pair of homologues, with one crossover results in one-half recombinant gametes)
- **Define** Expected No. recombinants as mapping function ($m = 0.5 \lambda$)
R.F. $r = 0.5(1 - e^{-2m})$ (form of Haldane's M.F.)
with inverse $m = -0.5 \ln(1 - 2r)$
converting an estimated R.F. to Haldane's map distance
- Thus, for locus order ABC
 $m_{AC} = m_{AB} + m_{BC}$ (since $m_{AB} = -0.5 \ln(1 - 2r_{AB})$) etc.
Substituting for each of these gives us the usual relationship between R.F.'s (no interference situation)
- Net Effect - **transform to straight line** m_{AC} vs m_{AB} or m_{BC}
- In practice - too simple; only applies to specific conditions; may not relate directly to physical distance -(common M.F. problem).

2

Example

RECOMBINANTS and MULTINOMIAL

- **Binomial** No. of recombinant gametes, produced by a heterozygous parent for a 2-locus model, with $\theta = P\{\text{gamete recombinant}\}$ (=R.F.)

So for r recombinants in sample of n

$$P\{X = r\} = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

- **Multinomial** 3-locus model (A,B,C) - 4 possible classes of gametes (non-recombinants, AB recombinants, BC recombinants and double recombinants loci ABC). Joint probability distribution for r.v.'s counting number in each class

$$P\{X_1 = a, X_2 = b, X_3 = c, X_4 = d\} = \frac{n!}{a!b!c!d!} P_1^a P_2^b P_3^c P_4^d$$

where $a+b+c+d=n$ and P_1, P_2, P_3, P_4 are probabilities of observing a member of each of 4 classes respectively

3

Central Limit Theorem

Suppose that we repeatedly draw random samples of size n (with replacement) from a distribution with mean μ and variance σ^2 . Let $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$ be the collection of sample averages and let

$$\bar{x}_i = \frac{\bar{x}_i - \mu}{\sigma/\sqrt{n}} \quad i = 1, 2, \dots$$

where the collection $\bar{x}_1, \bar{x}_2, \dots$ is called the **sampling distribution of means**.

Central Limit Theorem.

If X_1, X_2, \dots, X_n are a random sample of r.v. X , (mean μ , variance σ^2), then, in the limit, as $n \rightarrow \infty$, the sampling distribution of means has a Standard Normal distribution, $N(0,1)$

4

Probabilities for sampling distribution

- **Statements** $P\left\{a < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < b\right\} \cong P\{a < U < b\}$ for large n

U = standardized Normal deviate

- and, in particular $P\{|\bar{x} - \mu| < r\} = P\{-r < \bar{x} - \mu < r\}$

$$= P\left\{\frac{-r}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{r}{\sigma/\sqrt{n}}\right\}$$

$$= F\left(\frac{r}{\sigma/\sqrt{n}}\right) - F\left(\frac{-r}{\sigma/\sqrt{n}}\right)$$

- In general, closer r.v. X to Normal, the faster the approximation approaches U. Generally $n \geq 30 \Rightarrow$ "Large sample" theory

5

Probability Statements

- If X and Y independent Binomially distributed r.v.'s parameters n, p and m, p respectively, then $X+Y \sim B(n+m, p)$ - (show e.g. by m.g.f.'s)

- So, $Y = X_1 + X_2 + \dots + X_n \sim B(n, p)$ for the IID $X \sim B(1, p)$.
- Since we know $\mu_y = np$, $\sigma_y = \sqrt{npq}$ and, clearly $Y = n\bar{x}$ then

$$\frac{\bar{x} - \mu_x}{\sigma_x} = \frac{\frac{Y}{n} - \frac{\mu_y}{n}}{\frac{\sigma_y}{n}} = \frac{Y - np}{\sqrt{npq}} \rightarrow N(0,1) \text{ as } n \rightarrow \infty$$

and, further $U = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$ sampling distribution of a proportion

6

Probability Statements- C.L.T. and Approximation summary

- General form of theorem - an infinite sequence of independent r.v.'s, with means, variances as before, then approximation $\rightarrow U$ for n large enough. **Note:** No condition on **form** of distribution of X 's (raw data)
- **Strictly** - for approximations of discrete distributions, can improve by considering **correction for continuity**

e.g.

$$U \cong \frac{X - \lambda \pm 0.5}{\sqrt{\lambda}} \quad \text{Poisson, parameter } \lambda$$

$$U \cong \frac{(x/n) \pm 0.5 - p}{\sqrt{pq/n}} \quad x = \text{No. in sample, proportion} = \hat{p}$$

7

Generalising Sampling Distn. concept

- For the sampling distribution of any statistic. We say that a sample characteristic is an **unbiased estimator** of the parent population characteristic, if the **mean** of the corresponding sampling distribution is equal to the parent characteristic.

- **Lemma.** The sample average (proportion) is an unbiased estimator of the parent average (proportion): $E\{\bar{x}\} = \mu \quad E\{\hat{p}\} = P$

Recall: Sampling **without replacement** from **finite population - Hypergeometric**. The quantity $\sqrt{[(N-n)/(N-1)]}$ is called the **finite population correction (fpc)**. If the parent population is infinite or if sampling **with replacement** the fpc = 1.

- **Lemma.** $E[s] = S \times \text{fpc}$.

8

LIKELIHOOD - DEFINITIONS

- Suppose X can take a set of values x_1, x_2, \dots with

$$L(\theta) \propto P\{X = x|\theta\}$$

where θ is a vector of parameters affecting observed x 's

- e.g. $X \sim N(\mu, \sigma^2)$. So can say something about $P\{X\}$ if we know say $\theta = (\mu, \sigma^2)$
- **But** not usually case, i.e. observe x 's, knowing nothing of θ
- Assuming x 's a random sample size n from a known distribution, then likelihood for θ

$$L(\theta) = L(\theta|x_1, x_2, \dots, x_n) = \prod_{i=1}^n L(x_i|\theta)$$

- Finding most likely θ for given data is equivalent to Maximising the Likelihood function.
where M.L.E. is $\hat{\theta}$

9

LIKELIHOOD - SCORE and INFO. CONTENT

- The **Log-likelihood** is a **support function** $[S(\theta)]$ evaluated at point, θ' say
- Support function for any other point, say θ'' can be obtained approx., using the Taylor expansion
$$S(\theta'') = S(\theta') + (\theta'' - \theta') \frac{d[S(\theta')]}{d\theta} + \frac{1}{2} (\theta'' - \theta')^2 \frac{d^2[S(\theta')]}{d\theta^2} + \dots$$

and this is the basis of the Newton-Raphson iteration for the M.L.E.
- **SCORE** = first derivative of support function w.r.t. the parameter
$$= \frac{d[S(\theta)]}{d\theta} \quad \text{or, numerically, } \frac{S(\theta + \Delta) - S(\theta)}{\Delta}$$

- **INFORMATION CONTENT** evaluated at (i) arbitrary point = **Observed Info.** (ii) support function maximum = **Expected Info.**

$$I(\theta) = E_{\theta} \left[\left[\frac{\partial}{\partial \theta} \log L(\theta/x) \right]^2 \right] = -E_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \log L(\theta/x) \right]$$

10

Example - Binomial variable

(e.g. use of Score, Expected Info. Content to determine type of mapping population and sample size for genomics experiments)

Likelihood function

$$L(\theta) = L(n, p) = P\{X = x/n, p\} = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Log-likelihood

$$\text{Log}\{L(\theta)\} = \text{Log} \binom{n}{x} + x \text{Log} \theta + (n-x) \text{Log}(1-\theta)$$

Assuming n constant, then first term invariant w.r.t. $p = S(\theta)$ at point p

$$\text{Log}[L(p)] = x \text{Log} p + (n-x) \text{Log}(1-p)$$

Maximising w.r.t. p gives M.L.E. $\hat{\theta} = \hat{p} = \frac{x}{n}$ with **SCORE** $= \frac{x}{\theta} - \frac{n-x}{1-\theta}$

11

Bayesian Estimation- in context

- **Parametric Estimation** - in the "classical approach" $f(x, \theta)$ for a r.v. X of density $f(x)$, with θ the unknown parameter indicates the dependency of the distribution on the parameter to be estimated.
- **Bayesian Estimation** - θ is a random variable, so appropriate to consider the density as conditional and write $f(x|\theta)$
Given a random sample X_1, X_2, \dots, X_n the sample random variables can be considered jointly distributed with parameter r.v. θ . So, joint pdf

$$f_{X_1, X_2, \dots, X_n, \theta}(x_1, x_2, \dots, x_n, \theta)$$

- Objective - to form an estimator that gives a value of θ dependent on observations of the sample random variables. Thus conditional density of θ given X_1, X_2, \dots, X_n also plays a role. This is the **posterior density**

12

Bayes - contd.

- **Posterior Density** $f(\theta|x_1, x_2, \dots, x_n)$

- **Relationship - prior and posterior:**

$$f(\theta|x_1, x_2, \dots, x_n) = \frac{\pi(\theta) \prod_{k=1}^n f(x_k|\theta)}{\int_{-\infty}^{\infty} \pi(\theta) \prod_{k=1}^n f(x_k|\theta) d\theta}$$

where $\pi(\theta)$ prior density of θ

- **Value:** Close to MLE for large n , or for small n if sample values compatible with the prior distribution, strong sample basis, simpler to calculate.