

Lecture 1: Basic Statistical Tools

Bruce Walsh lecture notes
Uppsala EQG 2012 course
version 28 Jan 2012

Discrete Random Variables

A **random variable** (RV) = outcome (**realization**) not a set value, but rather drawn from some probability distribution

A **discrete** RV x --- takes on values X_1, X_2, \dots, X_k

Probability distribution: $P_i = \Pr(x = X_i)$

Probabilities are non-negative and sum to one $P_i \geq 0, \quad \sum P_i = 1$

Example: **Binominal** random variable. Let p = prob of a success.

Prob (k successes in n trials) = $n! / [(n-k)! k!] p^k (1-p)^{n-k}$

Example: **Poisson** random variable. Let λ = expected number of successes. Prob (k successes) = $\lambda^k \exp(-\lambda) / k!$

Continuous Random Variables

A **continuous** RV x can take on any possible value in some interval (or set of intervals). The probability distribution is defined by the **probability density function**, $p(x)$

$$p(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(x) dx = 1$$

Prob is area under
the curve

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x) dx$$

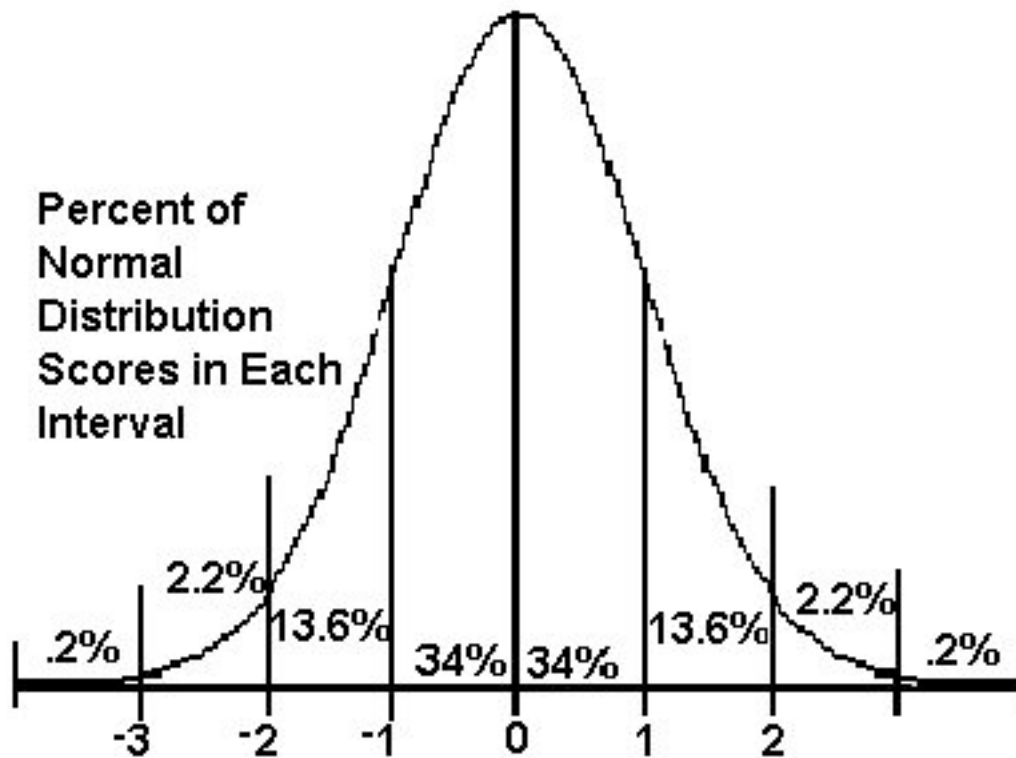
Finally, the **cdf**, or **cumulative probability function**, is defined as $\text{cdf}(z) = \text{Pr}(x \leq z)$

$$\text{cdf}(x) = \int_{-\infty}^x p(x) dx$$

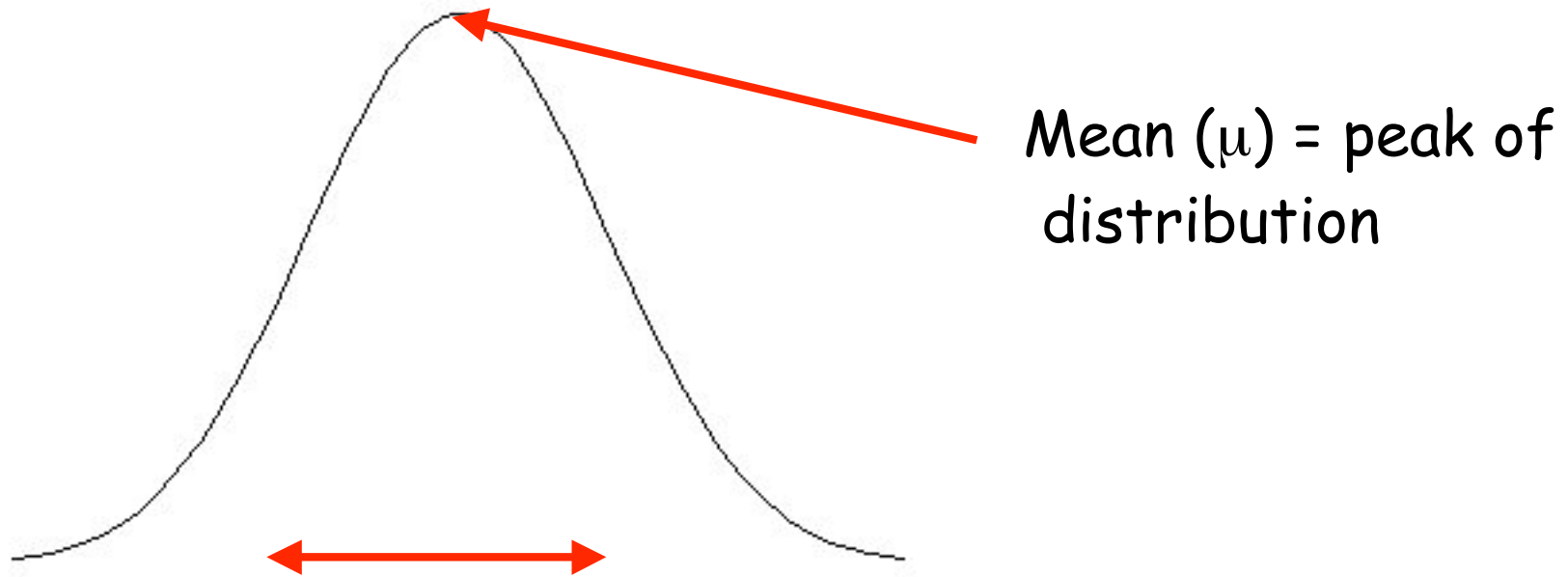
Example: The normal (or Gaussian) distribution

$$\phi(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Mean μ , variance σ^2



Unit normal
(mean 0, variance 1)



The variance is a measure of spread about the mean. The smaller σ^2 , the narrower the distribution about the mean

If $x \sim N(0, 1)$, $y \sim N(\mu, \sigma^2)$, then

$$\sigma \cdot (x + \mu) \sim N(\mu, \sigma^2)$$

$$\frac{y - \mu}{\sigma} \sim N(0, 1)$$

Joint and Conditional Probabilities

The probability for a pair (x,y) of random variables is specified by the **joint probability density function**, $p(x,y)$

$$P(y_1 \leq y \leq y_2, x_1 \leq x \leq x_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} p(x, y) dx dy$$

The **marginal density** of x , $p(x)$

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy$$

Joint and Conditional Probabilities

$p(y|x)$, the conditional density y given x

$$P(y_1 \leq y \leq y_2 | x) = \int_{y_1}^{y_2} p(y | x) dy$$

Relationships among $p(x)$, $p(x,y)$, $p(y|x)$

x and y are said to be independent if $p(x,y) = p(x)p(y)$

$$p(x,y) = p(y|x)p(x), \quad \text{hence} \quad p(y|x) = \frac{p(x,y)}{p(x)}$$

Note that $p(y|x) = p(y)$ if x and y are independent

Bayes' Theorem

Suppose an unobservable RV takes on values $b_1 \dots b_n$

Suppose that we observe the outcome A of an RV correlated with b . What can we say about b given A ?

Bayes' theorem:

$$\Pr(b_j | A) = \frac{\Pr(b_j) \Pr(A | b_j)}{\Pr(A)} = \frac{\Pr(b_j) \Pr(A | b_j)}{\sum_{i=1}^n \Pr(b_i) \Pr(A | b_i)}$$

A typical application in genetics is that A is some phenotype and b indexes some underlying (but unknown) genotype

Example: BRCA1/2 & Breast cancer

- NCI statistics:
 - 12% is lifetime risk of breast cancer in females
 - 60% is lifetime risk if carry BRCA 1 or 2 mutation
 - One estimate of BRCA 1 or 2 allele frequency is around 2.3%.
 - Question: Given a patient has breast cancer, what is the chance that she has a BRCA 1 or BRCA 2 mutation?

- Here
 - Event B = has a BRCA mutation
 - Event A = has breast cancer
- Bayes: $\Pr(B|A) = \Pr(A|B) * \Pr(B) / \Pr(A)$
 - $\Pr(A) = 0.12$
 - $\Pr(B) = 0.023$
 - $\Pr(A|B) = 0.60$
 - Hence, $\Pr(\text{BRCA}|\text{Breast cancer}) = [0.60 * 0.023] / 0.12 = 0.115$
- Hence, for the assumed BRCA frequency (2.3%), 11.5% of all patients with breast cancer have a BRCA mutation

Second example: Suppose height > 70. What is
The probability individual is QQ, Qq, qq?

Genotype	QQ	Qq	qq
Freq(genotype)	0.5	0.3	0.2
Pr(height >70 genotype)	0.3	0.6	0.9

$$\Pr(\text{height} > 70) = 0.3 \cdot 0.5 + 0.6 \cdot 0.3 + 0.9 \cdot 0.2 = 0.51$$

$$\Pr(\text{QQ} \mid \text{height} > 70) = \frac{\Pr(\text{QQ}) * \Pr(\text{height} > 70 \mid \text{QQ})}{\Pr(\text{height} > 70)}$$

$$= 0.5 \cdot 0.3 / 0.51 = 0.294$$

Expectations of Random Variables

The **expected value**, $E[f(x)]$, of some function x of the random variable x is just the average value of that function

$$E[f(x)] = \sum_i \Pr(x = X_i) f(X_i) \quad \times \text{ discrete}$$

$$E[f(x)] = \int_{-\infty}^{+\infty} f(x) p(x) dx \quad \times \text{ continuous}$$

$E[x]$ = the (arithmetic) mean, μ , of a random variable x

$$E(x) = \mu = \int_{-\infty}^{+\infty} x p(x) dx$$

Expectations of Random Variables

$E[(x - \mu)^2] = \sigma^2$, the variance of x

$$E[(x - \mu)^2] = \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx$$

More generally, the r th moment about the mean is given by $E[(x - \mu)^r]$ $r = 2$: variance. $r = 3$: skew

$r = 4$: (scaled) kurtosis

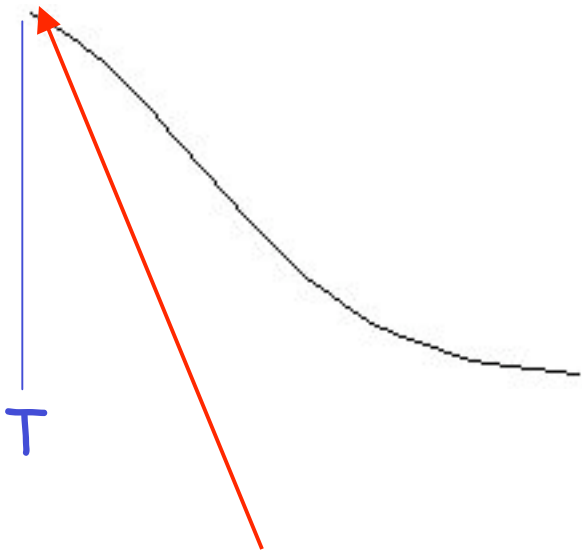
Useful properties of expectations

$$E[g(x) + f(y)] = E[g(x)] + E[f(y)]$$

$$E(cx) = cE(x)$$

The truncated normal

Only consider values of T or above in a normal



Here p_T is the height of the normal at the truncation point,

Density function = $p(x \mid x > T)$

$$\frac{p(z)}{\Pr(z > T)} = \frac{p(z)}{\int_T^\infty p(z) dz}$$

Mean of truncated distribution

$$E[z \mid z > T] = \int_T^\infty \frac{z p(z)}{\pi_T} dz = \mu + \frac{\sigma \cdot p_T}{\pi_T}$$

Let $\pi_T = \Pr(z > T)$

$$p_T = (2\pi)^{-1/2} \exp \left[-\frac{(T - \mu)^2}{2\sigma^2} \right]$$

The truncated normal

Variance

$$\left[1 + \frac{p_T \cdot (z - \mu)/\sigma}{\pi_T} - \left(\frac{p_T}{\pi_T} \right)^2 \right] \sigma^2$$

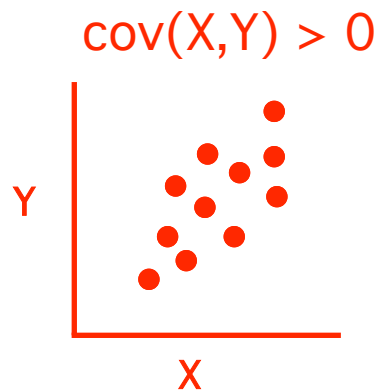
p_T = height of normal at T

π_T = area of curve to right of T

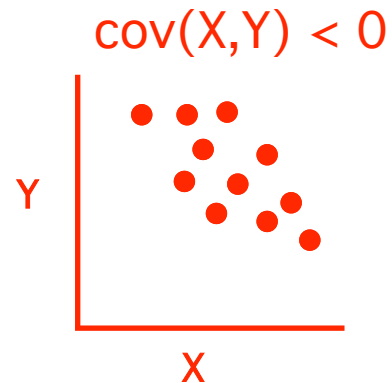
Covariances

- $\text{Cov}(x,y) = E [(x-\mu_x)(y-\mu_y)]$
 - $= E [x*y] - E[x]*E[y]$

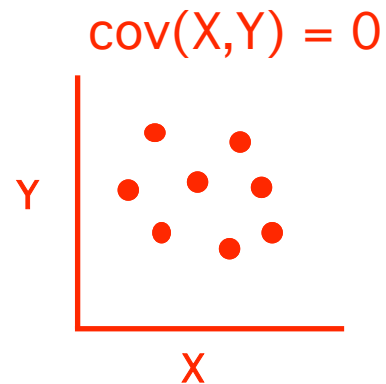
$\text{Cov}(x,y) > 0$, positive (linear) association between x & y



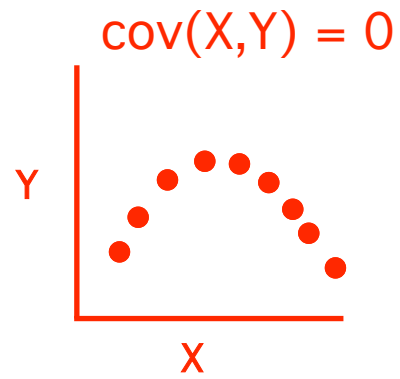
$\text{Cov}(x,y) < 0$, negative (linear) association between x & y



$\text{Cov}(x,y) = 0$, no *linear* association between x & y



$\text{Cov}(x,y) = 0$ DOES NOT imply no association



Correlation

Cov = 10 tells us nothing about the strength of an association

What is needed is an absolute measure of association

This is provided by the **correlation**, $r(x,y)$

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x) Var(y)}}$$

$r = 1$ implies a perfect (positive) linear association

$r = -1$ implies a perfect (negative) linear association

Useful Properties of Variances and Covariances

- Symmetry, $\text{Cov}(x,y) = \text{Cov}(y,x)$
- The covariance of a variable with itself is the variance, $\text{Cov}(x,x) = \text{Var}(x)$
- If a is a constant, then
 - $\text{Cov}(ax,y) = a \text{Cov}(x,y)$
- $\text{Var}(a x) = a^2 \text{Var}(x)$.
 - $\text{Var}(ax) = \text{Cov}(ax,ax) = a^2 \text{Cov}(x,x) = a^2 \text{Var}(x)$
- $\text{Cov}(x+y,z) = \text{Cov}(x,z) + \text{Cov}(y,z)$

More generally

$$\text{Cov} \left(\sum_{i=1}^n x_i, \sum_{j=1}^m y_j \right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(x_i, y_j)$$

$$\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y) + 2\text{Cov}(x, y)$$

Hence, the variance of a sum equals the sum of the Variances ONLY when the elements are uncorrelated

Question: What is $\text{Var}(x-y)$?

Regressions

Consider the best (linear) predictor of y given we know x

$$\hat{y} = \bar{y} + b_{y|x}(x - \bar{x})$$

The slope of this linear regression is a function of Cov,

$$b_{y|x} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

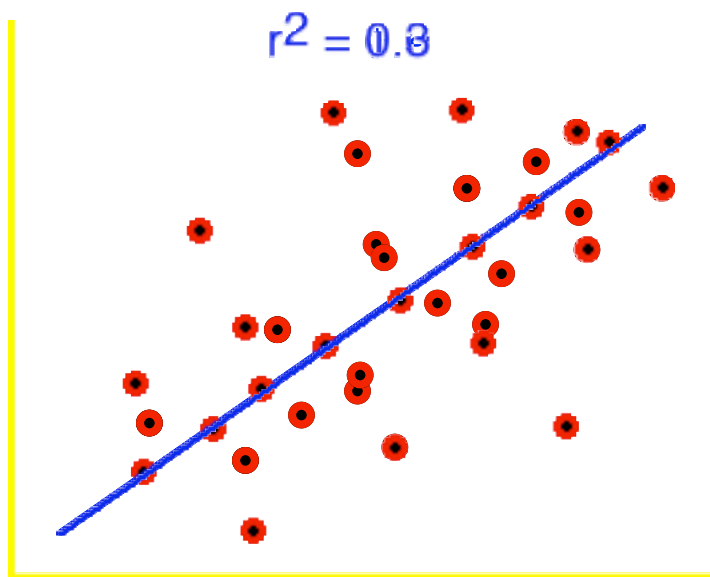
The fraction of the variation in y accounted for by knowing x , i.e., $\text{Var}(\hat{y} - y)$, is r^2

Relationship between the correlation and the regression slope:

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} = b_{y|x} \sqrt{\frac{Var(x)}{Var(y)}}$$

If $Var(x) = Var(y)$, then $b_{y|x} = b_{x|y} = r(x, y)$

In this case, the fraction of variation accounted for by the regression is b^2



Properties of Least-squares Regressions

The slope and intercept obtained by least-squares:
minimize the sum of squared residuals:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$

- The average value of the residual is zero
- The LS solution maximizes the amount of variation in y that can be explained by a linear regression on x
- Fraction of variance in y accounted by the regression is r^2
- The residual errors around the least-squares regression are uncorrelated with the predictor variable x
- **Homoscedastic** vs. **heteroscedastic** residual variances²⁵

Different methods of analysis

- Parameters of these various models can be estimated in a number of frameworks
- **Method of moments**
 - **Very little assumptions about the underlying distribution.** Typically, the mean of some statistic has an expected value of the parameter
 - Example: Estimate of the mean μ given by the sample mean, \bar{x} , as $E(\bar{x}) = \mu$.
 - While estimation does not require distribution assumptions, **confidence intervals and hypothesis testing do**
- **Distribution-based estimation**
 - **The explicit form of the distribution used**

Distribution-based estimation

- Maximum likelihood estimation
 - MLE
 - REML
 - More in Lynch & Walsh (book) Appendix 3
- Bayesian
 - Marginal posteriors
 - Conjugating priors
 - MCMC/Gibbs sampling
 - More in Walsh & Lynch (online chapters = Vol 2) Appendices 2,3

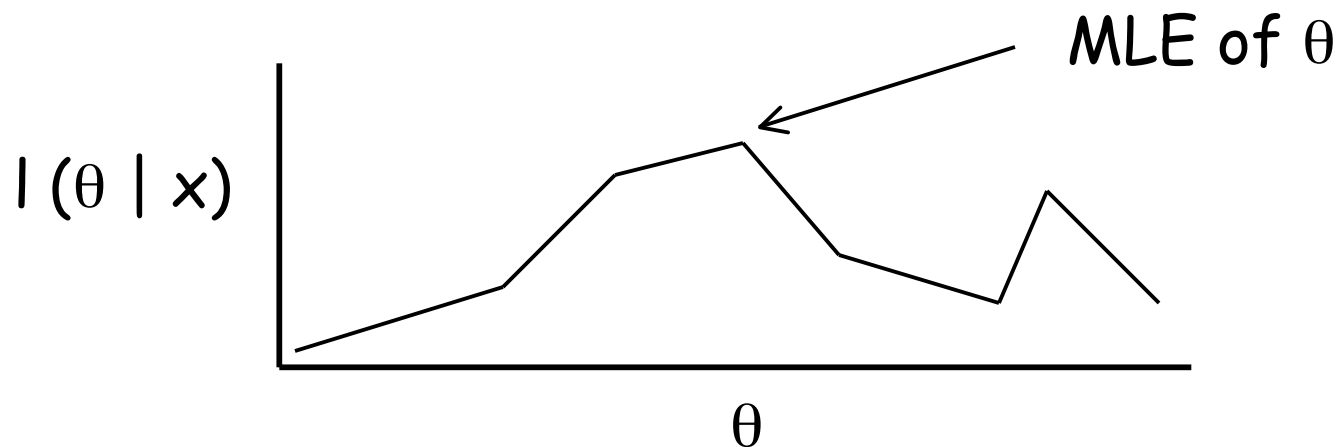
Maximum Likelihood

$p(x_1, \dots, x_n \mid \theta)$ = density of the observed data (x_1, \dots, x_n) given the (unknown) distribution parameter(s) θ

Fisher suggested the method of maximum likelihood --- given the data (x_1, \dots, x_n) find the value(s) of θ that **maximize** $p(x_1, \dots, x_n \mid \theta)$

We usually express $p(x_1, \dots, x_n \mid \theta)$ as a **likelihood function** $l(\theta \mid x_1, \dots, x_n)$ to remind us that it is dependent on the observed data

The **Maximum Likelihood Estimator (MLE)** of θ are the value(s) that maximize the likelihood function l given the observed data x_1, \dots, x_n .



This is formalized by looking at the **log-likelihood surface**, $L = \ln [l(\theta | x)]$. Since \ln is a monotonic function, the value of θ that maximizes l also maximizes L .

The curvature of the likelihood surface in the neighborhood of the MLE informs us as to the precision of the estimator. A narrow peak = high precision. A broad peak = lower precision.

$$\text{Var}(\text{MLE}) = -1 / \frac{\partial^2 L(\mu | z)}{\partial \mu^2}$$

The larger the curvature, the smaller the variance

Likelihood Ratio tests

Hypothesis testing in the ML frameworks occurs through **likelihood-ratio (LR) tests**

$$LR = 2 \ln \left(\frac{\ell(\hat{\Theta}_r | \mathbf{z})}{\ell(\hat{\Theta} | \mathbf{z})} \right) = 2 \left[L(\hat{\Theta}_r | \mathbf{z}) - L(\hat{\Theta} | \mathbf{z}) \right]$$

$\hat{\Theta}_r$ is the MLE under the restricted conditions (some parameters specified, e.g., var = 1)

$\hat{\Theta}$ is the MLE under the unrestricted conditions (no parameters specified)

For large sample sizes (generally) LR approaches a Chi-square distribution with r df (r = number of parameters assigned fixed values under null)

Bayesian Statistics

An extension of likelihood is Bayesian statistics

Instead of simply estimating a point estimate (e.g., the MLE), the goal is to **estimate the entire distribution** for the unknown parameter θ given the data x

$$p(\theta | x) = C * l(x | \theta) p(\theta)$$

$p(\theta | x)$ is the **posterior distribution** for θ given the data x

$l(x | \theta)$ is just the likelihood function

$p(\theta)$ is the **prior distribution** on θ .

Bayesian Statistics

Why Bayesian?

- Exact for any sample size
- Marginal posteriors
- Efficient use of any prior information
- MCMC (such as Gibbs sampling) methods

Priors quantify the strength of any prior information. Often these are taken to be diffuse (with a high variance), so prior weights on θ spread over a wide range of possible values.

Marginal posteriors

- Often times we are interested in a particular set of parameters (say some subset of the fixed effects). However, we also have to estimate all of the other parameters.
- How do uncertainties in these **nuisance parameters** factor into the uncertainty in the parameters of interest?
- A Bayesian marginal posterior takes this into account by integrating the full posterior over the nuisance parameters
- While this sounds complicated, easy to do with MCMC (Markov Chain Monte Carlo)

Conjugating priors

For any particular likelihood, we can often find a **conjugating prior**, such that **the product of the likelihood and the prior returns a known distribution.**

Example: For the mean μ in a normal, taking the prior on the mean to also be normal returns a posterior for μ that is normal.

Example: For the variance σ^2 in a normal, taking the prior on the variance to an inverse chi-square distribution returns a posterior for σ^2 that is also an inverse chi-square (details in WL Appendix 2).

A normal prior on the mean with mean μ_0 and variance σ_0^2 (larger σ_0^2 , more diffuse the prior)

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

If the likelihood for the mean is a normal distribution, the resulting posterior is also normal, with

$$\sigma_*^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_* = \sigma_*^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right)$$

Note that if σ_0^2 is large, the mean of the posterior is very close to the sample mean.

If x follows a Chi-square distribution, then $1/x$ follows an **inverse chi-square distribution**.

The **scaled inverse chi-square distribution** is more typically used, where

$$p(x | n) \propto x^{-(n/2+1)} e^{-\sigma_0^2/(2x)}$$

The scaled inverse chi-square has two parameters, allowing more control over the mean and variance of the prior

Likelihood	Conjugate prior
Normal	
μ unknown, σ^2 known	Normal
μ known, σ^2 unknown	Inverse Chi-Square
Multivariate Normal	
μ unknown, \mathbf{V} known	Multivariate Normal
μ known, \mathbf{V} unknown	Inverse Wishart

MCMC

Analytic expressions for posteriors can be complicated, but the method of **MCMC** (**Markov Chain Monte Carlo**) is a general approach to simulating draws for just about any distribution (details in WL Appendix 3).

Generating several thousand such draws from the posterior returns an *empirical distribution* that we can use.

For example, we can compute a 95% credible interval, the region of the distribution that containing 95% of the probability.

Gibbs Sampling

- A very powerful version of MCMC is the [Gibbs Sampler](#)
- Assume we are sampling from a vector of parameters, but that the marginal distribution of each parameter is known
- For example, given a current value for all the fixed effects (but one, say β_1) and the variances, conditioning on these values the distribution of β_1 is a normal, whose parameters are now functions of the current values of the other parameters. A random draw is then generated from this distribution.
- Likewise, conditioning on all the fixed effects and all variances but one, the distribution of this variance is an inverse chi-square

When more than two variables are involved, the sampler is extended in the obvious fashion. In particular, the value of the k th variable is drawn from the distribution $p(\theta^{(k)} | \Theta^{(-k)})$ where $\Theta^{(-k)}$ denotes a vector containing all of the variables but k . Thus, during the i th iteration of the sample, to obtain the value of $\theta_i^{(k)}$ we draw from the distribution

$$\theta_i^{(k)} \sim p(\theta^{(k)} | \theta^{(1)} = \theta_i^{(1)}, \dots, \theta^{(k-1)} = \theta_i^{(k-1)}, \theta^{(k+1)} = \theta_{i-1}^{(k+1)}, \dots, \theta^{(n)} = \theta_{i-1}^{(n)})$$

For example, if there are four variables, (w, x, y, z) , the sampler becomes

$$w_i \sim p(w | x = x_{i-1}, y = y_{i-1}, z = z_{i-1})$$

$$x_i \sim p(x | w = w_i, y = y_{i-1}, z = z_{i-1})$$

$$y_i \sim p(y | w = w_i, x = x_i, z = z_{i-1})$$

$$z_i \sim p(z | w = w_i, x = x_i, y = y_i)$$

This generates one cycle of the sampler. Using these new values, a second cycle is generated.

Full details in WL Appendix 3.