Lecture 3:
Basic Statistical Tools

Bruce Walsh lecture notes
Tucson Winter Institute
7 - 9 Jan 2013

Basic probability

- Events are possible outcomes from some random process
  - e.g., a genotype is *AA*, a phenotype is larger than 20
- Pr(E) denotes the probability of an event E
- Pr(E) is between zero and one
- The sum of the probabilities of all possible nonoverlapping events is one.
  - e.g, if the possible events are E_1 , ⋯ , E_k, then
  - $Pr(E_1) + \cdots + Pr(E_k) = 1$

The AND rule

- Consider two possible events, E_1 and E_2.
- If these are independent (knowledge that one has occurred does not change the probability of the second), then the joint probability $Pr(E_1,E_2)$, the Probability of E_1 AND E_2 is $Pr(E_1,E_2), = Pr(E_1)* Pr(E_2)$,
- Hence, with independence, AND = multiply
- Conditional probability is used when the events are NOT independent

Example

- Consider the cross AaBbCc X aaBbCc
  - What is the probability of an aabbcc offspring?
  - Assuming independent assortment (no linkage)
  - = Pr(aa | Aa x aa) * Pr(bb | Bb x Bb) * Pr(cc | Cc x Cc) = (1/2)(1/4)(1/4) = 1/32
- How many offspring do we need to score to have a 90% probability of seeing at least one?
  - Let p = 1/32. Prob(not seeing aabbcc in n offspring) = $(1-p)^n$.
  - Prob(at least one) = 0.9 implies Prob(none) = 0.1
  - $(1-p)^n = 0.1$, or n = log(0.1)/log(1-1/32) = 72.5

The OR rule

- Again, consider two possible events, E_1 and E_2.
- If these events are NONOVERLAPPING (they contain no common elements), then $Pr(E_1 \text{ or } E_2) = Pr(E_1) + Pr(E_2)$
- Hence, OR = add
- Example:
  - What is the probability that a genotype is *A*-, i.e., that is *AA* or *Aa*?
  - The events genotype = *AA* and genotype = *Aa* are nonoverlapping
  - Hence, Pr(*A*-) = Pr(*AA* or *Aa*) = Pr(*AA*) + Pr(*Aa*)

Conditional Probability

- It is ALWAYS true that
  - $Pr(A,B) = P(A|B)P(B) = P(B|A)P(A)$
  - $P(A|B)$ is the conditional probability of A given B
  - $P(A)$ is the marginal probability of A
  - $P(A,B)$ is the joint probability of A and B
  - If $P(A|B) = P(A)$ for all possible B values, then A and B are independent
- Note that
  - $P(A|B) = P(A,B)/P(B)$

Examples of Prob (cont)

- Recall that yellow peas (Y-) are dominant to green peas (gg). Consider the F_2 in a cross of YY x gg.
  - What is the probability of a yellow F_2 offspring?
    - Pr(yellow) = Pr(YY or Yg) = Pr(YY) + Pr(Yg) =1/4 + 1/2 = 3/4
  - What is the probability that a yellow F_2 offspring is a YY homozygote?
    - Pr(YY | F_2 Yellow) = Pr(YY and F_2 Yellow)/Pr(F_2 yellow) = (1/4)/(3/4) = 1/3.

Bayes' Theorem

Suppose an unobservable random variable (RV) takes on values b_1 .. 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:

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

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 patent 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(BRCA|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?

Suppose:

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

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

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

$$= 0.5*0.3 / 0.51 = 0.294$$

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 $X_1, X_2, \ldots X_k$

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: Suppose the probability of seeing no individuals of genotype AABB in our sample is 0.1. What is the probability of seeing at least one?
Pr(none) + Pr(at least one) = 1, hence Pr(at least one) = 1-Pr(none) = 0.9

The Binominal Distribution

- What is the expected number of successes in a series of n trails where the probability p of success is the same for each trail?
- This is given by the binominal distribution,
  - $\Pr(k \text{ successes} \mid n, p) = n!/[(n-k)!\, k!]\; p^k (1-p)^{n-k}$
- Example: Suppose p = 0.05 and n = 10. What is the probability of seeing EXACTLY one success?
  - $\Pr(k=1) = 10!/(9!*1!)\; 0.05^1\; 0.95^9 = 10*\; 0.05^1\; 0.95^9 = 0.315$
- What is the probability of seeing AT LEAST one success?
  - $\Pr(k > 0) = 1 - \Pr(k=0) = 1-(1-0.05)^{10} = 0.401$

The Poisson Distribution

- Given that the expected number of successes in our sample is λ, what is the probability that we see k successes?
- This is given by the Poisson distribution
  - $\Pr(k \text{ successes} \mid \lambda) = e^{-\lambda} \lambda^k / k!$
- Example: suppose $\lambda = 0.5$.
  - $\Pr(k = 1) = e^{-0.5} 0.5^1 / 1! = 0.303$
  - $\Pr(\text{at least one success}) = 1 - \Pr(k = 0) = 1 - e^{-0.5} = 0.393$
- Connection with binominal: $\lambda = n*p$
  - Can either use Poisson as an approximation or when the sample size n is not given

The geometric distribution

- Given success probability p per trail, how many failures k occur before the first success?
- This is a waiting-time (as opposed to a counting) problem, and is given by the geometric distribution
  - Pr(k failures before a success) = $(1-p)^k p$
  - Example: Suppose p = 0.05. What is the probability of AT LEAST one success in the first 10 trails?
  - = 1 - Pr(none in 1st 10) = $1-(1-p)^{10}$ = 0.401

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 (or pdf), p(x)

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

Prob is area under the curve

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

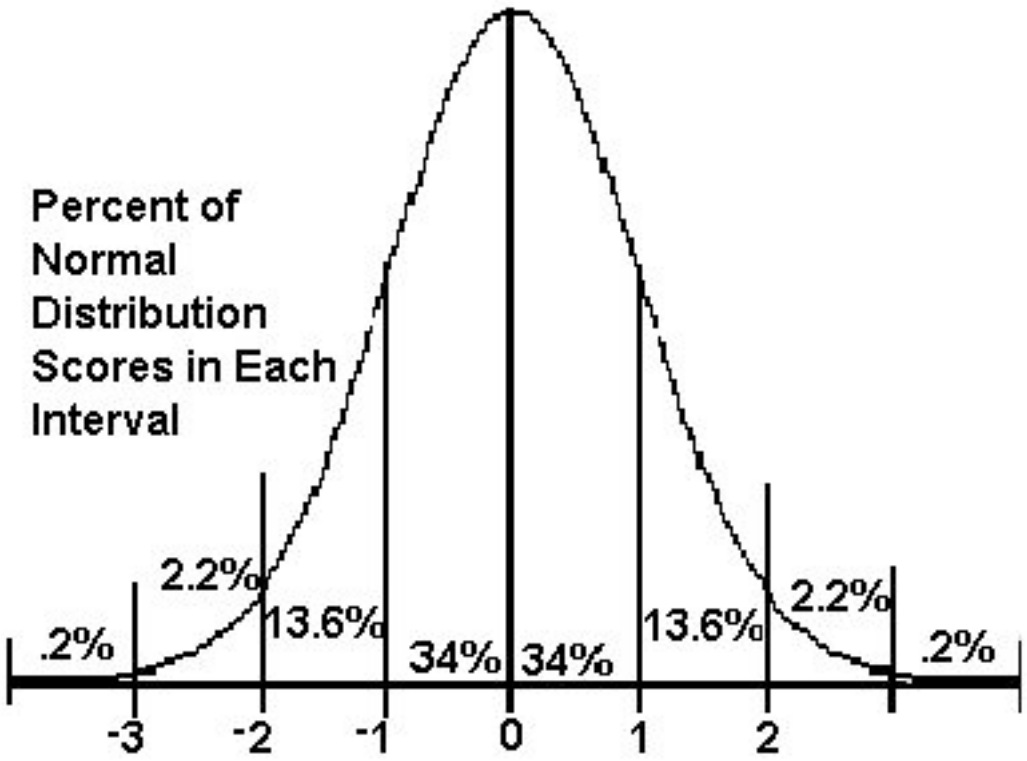
Finally, the cdf, or cumulative probability function, is defined as cdf(z) = Pr(x $\leq$ z)

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

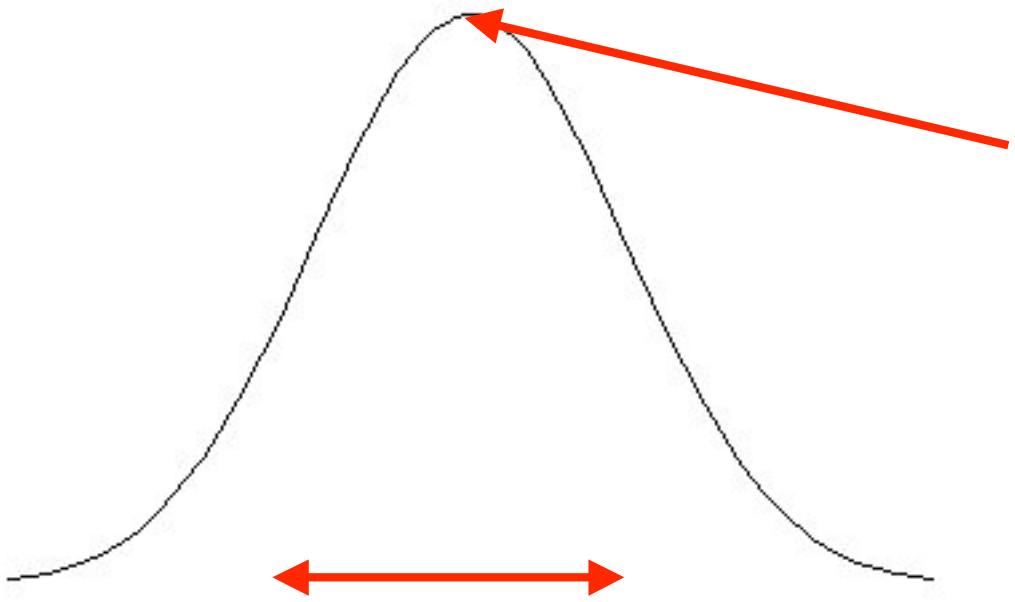
Example: The normal (or Gaussian) distribution

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

Mean μ, variance σ^2



Unit normal
(mean 0, variance 1)



Mean (μ) = peak of distribution

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

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

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

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

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 of y given x

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

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

Expectations of Random Variables

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

$$E[f(x)] = \sum_i \Pr(x = X_i) f(X_i)$$ x discrete

$$E[f(x)] = \int_{-\infty}^{+\infty} f(x) p(x) dx$$ x 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[$(x - \mu)^2$] = σ^2, the variance of x

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

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

r = 3: skew (value is zero for a normal)

r = 4: (scaled) kurtosis ($3\sigma^4$ for a normal)

Useful properties of expectations

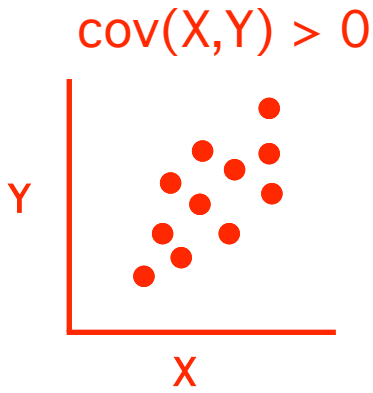
$$E[g(x) + f(y)] = E[g(x)] + E[f(y)]$$

$$E(c\,x) = c\,E(x)$$

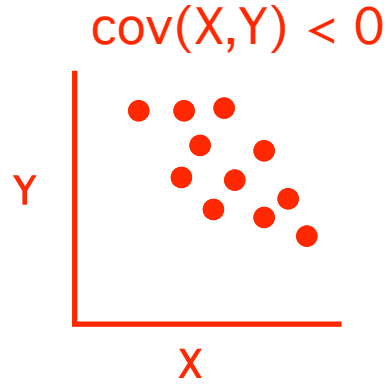
Covariances

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

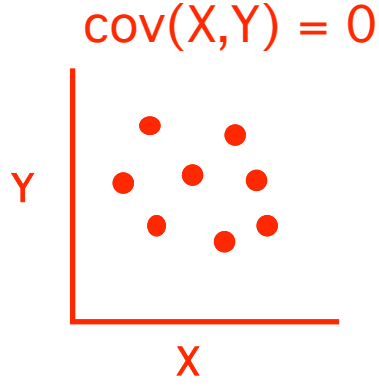
Cov(x,y) > 0, positive (linear) association between x & y



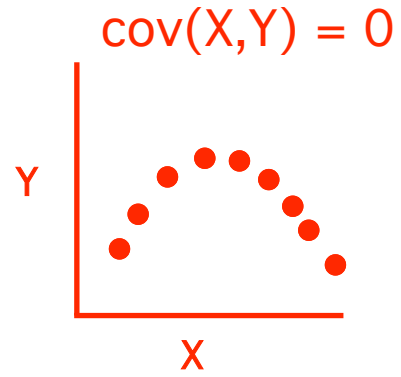
Cov(x,y) < 0, negative (linear) association between x & y



Cov(x,y) = 0, no *linear* association between x & y



Cov(x,y) = 0 DOES NOT imply <u>no association</u>



If x and y are independent, then cov(x,y) = 0

However, cov(x,y) = 0 DOES NOT imply that x and y are independent.

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

More generally

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

$$Var(x+y) = Var(x) + Var(y) + 2Cov(x,y)$$

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

Question: What is Var(x-y)?

Regressions

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

$$\widehat{y} = \overline{y} + b_{y\,|\,x}(x - \overline{x})$$

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

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

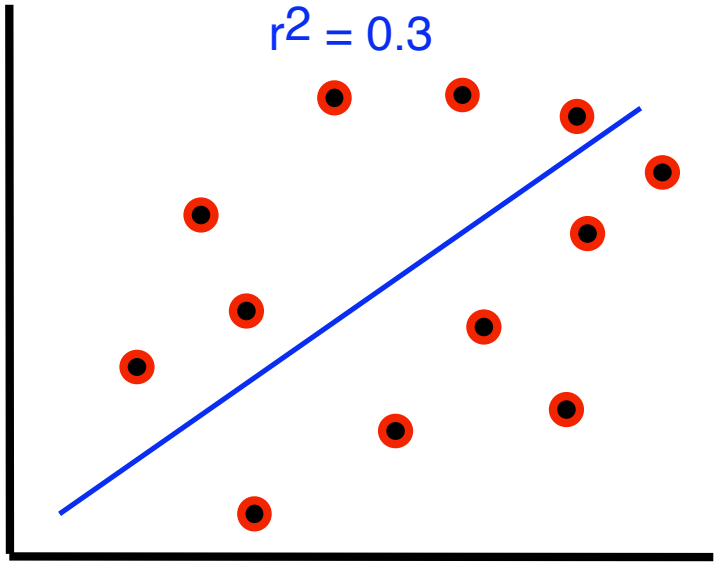
The fraction of the variation in y accounted for by knowing x, i.e,Var(yhat - y), is r^2

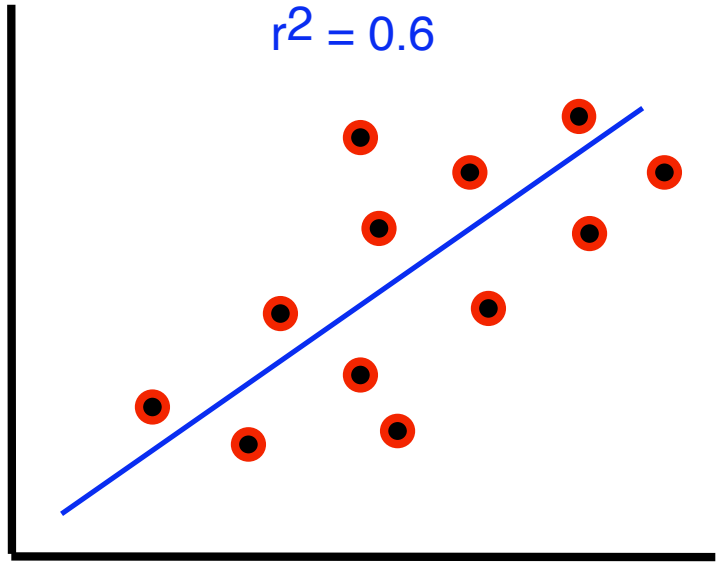
Relationship between the correlation and the regression slope:

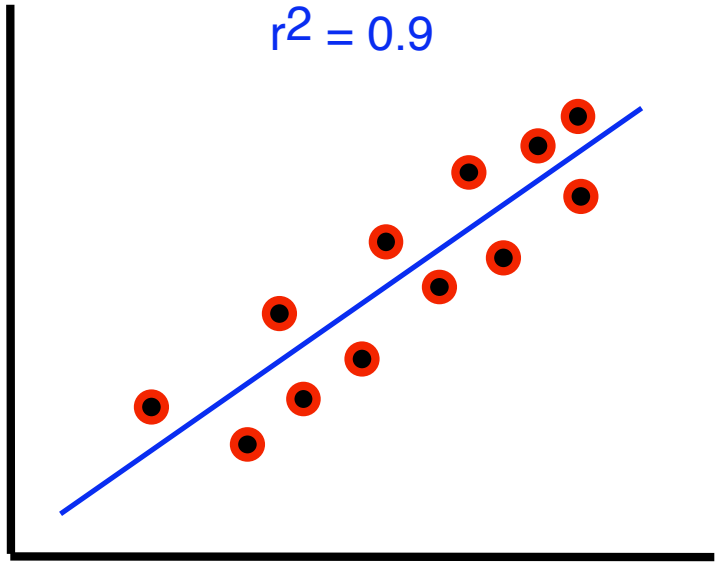
$$r(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = b_{y|x}\sqrt{\frac{Var(x)}{Var(y)}}$$

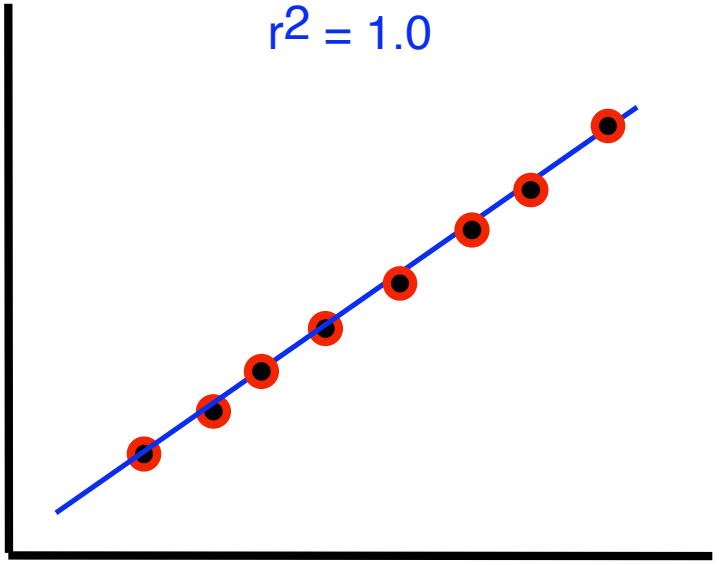
If Var(x) = Var(y), then $b_{y|x} = b_{x|y} = r(x,y)$

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


$r^2 = 0.3$


$r^2 = 0.6$


$r^2 = 0.9$


$r^2 = 1.0$

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 variances

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, xbar, as E(xbar) = μ.
  - 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
  - More in Walsh & Lynch (online chapters = Vol 2) Appendices 2,3

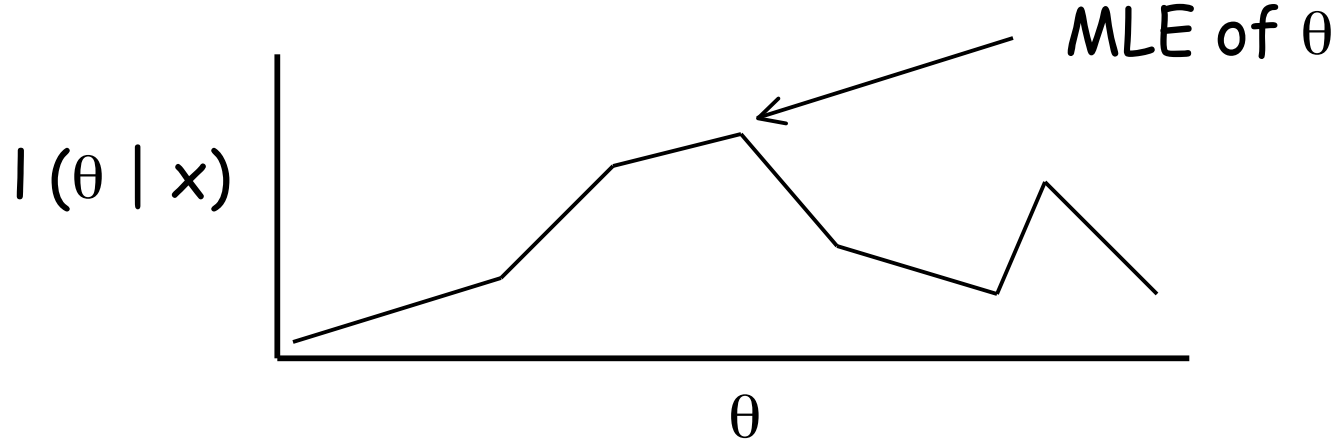
Maximum Likelihood

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

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

We usually express $p(x_1,\ldots, x_n \mid \theta)$ as a likelihood function $l(\theta \mid x_1,\ldots, x_n)$ 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 $x_1,\ldots, x_n$.



This is formalized by looking at the log-likelihood surface, L = ln [l (θ | 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 board peak = low precision

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

The larger the curvature, the smaller the variance

Likelihood Ratio tests

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

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

θ_r is the MLE under the restricted conditions (some parameters specified, e.g., var =1)

Θ_r 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 the estimate the entire distribution for the unknown parameter θ given the data x

$$p(\theta \mid x) = C * l(x \mid \theta)\, p(\theta)$$

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

$l(x \mid \theta)$ 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.

p values in Hypothesis testing

- The p value of a test statistic is the probability of seeing a value as large (or larger) under the null hypothesis
- For example, suppose you are assuming a random variable comes from a normal with mean zero and variance one.
  - The probability of seeing a value more extreme than 2 (i.e., greater than two or less than -2) is 0.0455, the p value associated with this value of the test statistic.

Significance and multiple comparisons

- One could either report a p value or have some criteria (i.e., any test with a p value less than 0.01) that declares a test to be significant (and hence a positive result)
  - p is the probability of a false positive, the probability of declaring a test under the null as being significant.
- The problem of multiple comparisons arises when a large number of tests are performed.
  - Suppose our significance threshold is p = 0.005, but 1000 tests are done. Under the null, we still expect 0.005*1000 = 5 significant tests
  - Bonferroni corrections are done by first setting a significance level for the entire COLLECTION of tests (say $\pi = 0.05$). To have this level experiment-wide control of false positives requires each test uses $p = \pi / n$
    - For n = 1000, an experiment-wide false positive rate (probability) of 0.05 declares significance only with the p value for a test is less than 0.05/1000 = 0.00005.

Power and Type I/II errors

- A Type I error is the probability of declaring a test to be significant when the null is true (a false positive)
- The power of a statistical test (a function of the sample size and the true parameters) is the probability of declaring a test to be significant when the null is false.
  - A Type II error occurs when we fail to declare a test significant when it is not from the null (i.e., a false negative)

FDR, the false discovery rate

- p is the probability of declaring a test under the null to be significant (the false-positive rate)
- When many tests are expected to be significant (i.e., looking for differences in expression over a large number of genes), a more appropriate measure is the false discovery rate (or FDR), the number of false positives among all tests declared to be significant.
  - Example: Suppose 1000 tests with a significant threshold of p = 0.05 is used. Expect 5 false positives, but suppose that 30 significant tests are found. Here the FDR = 5/30 = 0.167.
  - Hence, 16.7% of the positive tests are false positives