Central Limit Theorem

$\mu = \int xP(x)dx = p \times 1 + (1-p) \times 0 = p$
$\sigma^2 = \int (x-\mu)^2 p(x)dx = p(1-\mu)^2 + (1-p) \cdot (0-\mu)^2 = p(1-p)$
Binomial (n,p) $\approx$ Normal(np, $\sqrt{np(1-p)}$)
CLT applies to binomial because it's sum of Bernoulli's r.v.'s: N tries of an r.v. with values 1 (prob p) or 0 (prob 1-p)

Null hypothesis

Example of DNA
Model 1: $p_A = p_C = p_T = p_G = 0.25$
Model 2: $p_A = p_T,\ p_C = p_G$

Multinomial Model: At each position an i.i.d. choice of A,C,G,T with respective probabilities adding up to 1
Four multinomial model, e.g. choice of A vs. not A with some probability p_A
Binomial (n,p) $\approx$ Normal(np, $\sqrt{np(1-p)}$)

Model 1: all p's=0.25

$$\mu = 0.25N$$
$$\sigma = \sqrt{0.25 \times 0.75N}$$
$$t = \frac{n-\mu}{\sigma}$$
$$p = 2[1 - P_{Normal}(|t|)]$$

Model 2: A and T occur with identical probabilities, as do C and G.

$$\hat{p}_{AT} = \frac{1}{2}(n_A + n_T)/N$$
$$\hat{p}_{CG} = \frac{1}{2}(n_C + n_G)/N$$
$$n_A \sim Normal(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$n_T \sim Normal(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$\Rightarrow n_A - n_T \sim Normal(0, \sqrt{2N\hat{p}_{AT}(1-\hat{p}_{AT})})$$

The difference of two Normals is itself Normal; the variance of the sum is the sum of variances.

Bayesian hypothesis testing

Three bayesian criticisms of tail tests:

1. Their result depends on the choice of test or (more argumentatively) what was in the mind of the experimenter
   "Stopping rule paradoxes"
   Flipping coins, p=0.5. Result: 9 heads and 1 tail
   H0: a coin is fair with P(heads)=0.5
   Method 1:

$$p = \frac{1+10+1+10}{2^{10}} = 0.0214$$
$$p\ value < 0.01$$

Insignificant result

Method 2 : Protocol is to flip until a tail and record N.

$$H0: p(N) = 2^{-N+1}$$
$$p(\geq N) = 2^{-(N+1)}(1 + 1/2 + 1/4 + \dots) = 2^{-N}$$
$$p(\geq 9) = 2^{-9} = 0.00195 < 0.01$$

Significant result

Bayesian Approach

H_p: hypothesis that probability=p

$P(H_p)$ is the probability of the hypothesis

$$P(H_p|data) \propto P(data|H_p)P(H_p) \propto p^9(1-p)$$
$$P(H_p|data) = \frac{p^9(1-p)}{\int_0^1 p^9(1-p)\,dp}$$

Likelihood Ratio

$$\frac{P(H_{0.5}|data)}{P(H_{max}|data)} = \frac{0.1074}{4.2616} = 0.0252$$

Bayes tail probability

$$\int_0^{0.5} P(H_p|data)\,dp = 0.0059$$

Non-linear least square fits

Example of coin making machine

- Printing machine produces biased heads/tails with P(heads)=p.
- p(x) depends on the machine temperature x, as well as five parameters b_1, b_2, b_3, b_4, b_5
- n coins are tossed and binomial probability p is measured.
- The outcome is plotted as $2p - 0.4$ = $2n_{head}/n - 0.4$

  Model :

$$f(x) = 2p - 0.4 = b_1 \cdot exp(-b_2 x) + b_3 \cdot exp(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2})$$

Goal: Determine the parameters b_i

Data are collected at various temperatures x_i

$2n_{heads}/n - 0.4$ is measured to approximate $2p - 0.4$ from n coin tosses.

Weighted Nonlinear Least Squares Fitting = χ^2 fitting = Maximum Likelihood Estimation of Parameters (MLE) = Bayesian Parameter Estimation

$$y_i = y(x_i|b) + e_i$$
$$e_i \sim N(0, \sigma_i)$$
$$e \sim N(0, \sum)$$

b is the model, y_i is the suppposedly measured value plus error based on the model at different temperatures x_i.

$$P(b|y_i) \propto P(y_i|b)P(b)$$
$$\propto \Pi_i \ exp[-\frac{1}{2}(\frac{y_i - y(x_i|b)}{\sigma_i})^2]P(b)$$
$$\propto exp[-\frac{1}{2}\sum_i(\frac{y_i - y(x_i|b)}{\sigma_i})^2]P(b)$$
$$\propto exp[-\frac{1}{2}\chi^2(b)]P(b)$$

Maximize $P(b|y_i) \Rightarrow$ Find the parameter value that minimizes χ^2.
We can temporarily set prior P(b)=1, which means that all b models are equally likely.

$$\chi^2 = \sum_i(\frac{y_i - y(x_i|b)}{\sigma_i})^2$$

y_i is the actual measured value.

Posterior distribution of fitted parameters
Taylor expansion

$$-\frac{1}{2}\chi^2(b) \approx -\frac{1}{2}\chi^2_{min} - \frac{1}{2}(b - b_0)^T[\frac{1}{2}\frac{\partial^2\chi^2}{\partial b \partial b}](b - b_0)$$

By choosing the point of expansion at χ_{min}, the first deriative (second term in Taylor series) drops out.
Hessian Matrix (2nd derivative matrix)
Then,

$$P(b|y_i) \propto exp[-\frac{1}{2}(b - b_0)^T(\sum_b)^{-1}(b - b_0)]P(b)$$
$$\sum_b = [\frac{1}{2}\frac{\partial^2\chi^2}{\partial b \partial b}]^{-1} \Rightarrow Covariance\ (standard\ error)\ matrix$$

If Taylor Series converges rapidly and the prior $P(b)$ is uniform, then posterior distribution of b's is multivariate Normal.

Posterior and Prior

Bayes' theorem calculates the renormalized pointwise product of the prior and the likelihood function, to produce the *posterior probability distribution*, which is the conditional distribution of the uncertain quantity given the data.
Similarly, the **prior probability** of a random event or an uncertain proposition is the unconditional probability that is assigned before any relevant evidence is taken into account.

$$f : R^n \to R$$

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

For f(x,y):

$$f : R^2 \to R$$

$$H_{(f(x,y))} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix}$$

$$\frac{\partial^2 f}{\partial x \partial y} \approx \frac{1}{2h}(\frac{f_{++} - f_{-+}}{2h} - \frac{f_{+-} - f_{--}}{2h}) \\ = \frac{1}{4h^2}(f_{++} + f_{--} - f_{+-} - f_{-+})$$

where,

$$f_{++} = f(\vec{r} + h\hat{x} + h\hat{y}) \\ f_{+-} = f(\vec{r} + h\hat{x} - h\hat{y})$$

χ^2: "statistic" defined as the sum of the squares of n independent t-values

$$\chi^2 = \sum_i (\frac{x_i - \mu_i}{\sigma_i})^2, \ x_i \sim N(\mu_i, \sigma_i)$$

In this case, i ranges from 1 to 5.

$$\chi^2 \sim Chisquare(v), \ v > 0 \\ p(\chi^2) d\chi^2 = \frac{1}{2^{0.5v}\Gamma(0.5v)}(\chi^2)^{0.5v-1} exp(-0.5\chi^2) \ d\chi^2, \ \chi^2 > 0$$

where $p(\chi^2)$ is a probability density distribution function of χ^2.
Gamma function is,

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, dt, \qquad \Re(z) > 0 \, .$$

Case: v=1

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \Rightarrow x \sim N(0,1) \\ y = x^2 \\ p_Y(y) dy = 2 p_X(x) dx \\ p_Y(y) = y^{-1/2} p_X(y^{1/2}) \sim Chisquare(1)$$

Multivariate Normal Distributions

The multivariate normal distribution of a *k*-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)^{\mathrm{T}}$ can be written in the following notation:

$$\mathbf{X} \sim \mathbf{N}(\mu, \sum)$$

Generalizes Normal (Gaussian) to M-dimensions

$$N(x|\mu, \sum) = \frac{1}{(2\pi)^{M/2} det(\sum)^{1/2}} \, exp[-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)]$$

where mean is a M-vector, and covariance is a $M \times M$ matrix.
Components x_i of vector x are correlated random variables.

$$mean: \ \mu =< x >$$
$$covariance: \ \sum =< (x-\mu)(x-\mu)^T >$$

Simple example

$$p(x_1, x_2) = \frac{1}{\sqrt{2\pi}\sigma_1} exp[-\frac{1}{2}(x_1-\mu_1)^2/\sigma_1^2] \cdot \frac{1}{\sqrt{2\pi}\sigma_2} exp[-\frac{1}{2}(x_2-\mu_2)^2/\sigma_2^2]$$

where x_1, x_2 are two independent variables

Covariance matrix : can be applied to any set of random variables, not just multivariate normal.

$$Cov(x, y) =< (x-\bar{x})(y-\bar{y}) >$$
$$C = C_{ij} = Cov(x_i, x_j) =< (x_i - \bar{x_i})(x_j - \bar{x_j}) >$$

The diagonal elements are the variances of the individual variables
The variance of any linear combination of random variables is a quadratic form in **C**:

$$Var(\sum a_i x_i) =< \sum_i a_i(x_i - \bar{x_i}) \sum_j a_j(x_j - \bar{x_j}) >= \alpha^T C \alpha$$

Example of Coin toss
X=#heads, Y=#tails

$$X + Y = n$$
$$< X > + < Y >= n$$
$$X - E[X] = -Y + E[Y]$$
$$cov(X, Y) =< (X- < X >)(Y- < Y >) >$$

Linear correlation matrix

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

r is useful as "test for correlation".