Statistics 111

Probability and Statistical Theory

The axioms of probability are:

- $P(\Omega) = 1$
- $P(A) \geq 0$ for all events A
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \varnothing$

Other facts that follow include:

- $P(A^C) = 1 - P(A)$
- $P(\varnothing) = 0$
- $P(A) \leq P(B)$ if $A \subseteq B$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional Probability

Definition. A and B are independent if $P(A) = P(A|B)$ or, equivalently, if $P(A \cap B) = P(A)P(B)$.

Events may also be considered as sets of random variables.

Definition. X and Y are independent random variables if $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$, that is, $F_{XY}(x, y) = F_X(x)F_Y(y)$. This is equivalent to $f_{XY}(x, y) = f_X(x)f_Y(y)$ and $f_{X|Y}(x \mid y) = f_X(x)$. (If f_{XY} factors, then X and Y are independent.)

Definition. Let X and Y be discrete random variables. Then, we define $P(Y = y \mid X = x) = P(X = x, Y = y) / P(X = x)$. For continuous random variables, we define $f_{Y|X}(y \mid x) = f_{XY}(x, y) / f_X(x)$.

Definition. $X_1, \ldots, X_n$ are independent if $F_{X1, \ldots, Xn}(x_1, \ldots, x_n) = F_{X1}(x_1) \ldots F_{Xn}(x_n)$. Equivalently, $f_{X1, \ldots, Xn}(x_1, \ldots, x_n) = f_{X1}(x_1) \ldots f_{Xn}(x_n)$ for all $x_1, \ldots, x_n \in \mathbf{R}$.

Proposition. Let X, Y be continuous random variables. Them $f_{XY}(x, y)$ $f_X(x)f_Y(y)$ if and only if $f_{X|Y}(x \mid y) = f_X(x)$.

Distributions of Functions of Random Variables

Let X and Y be independent random variables. Suppose $S = X + Y$. Then $f_S(s) = \int_{-\infty}^{\infty} f_x(s - y)f_Y(y)\, dy$. Suppose $S = X - Y$. Then, $f_S(s) = \int_{-\infty}^{\infty} f_x(s + y)f_Y(-y)\, dy$

Suppose $Z = XY$. Then, $h(z) = \int_{-\infty}^{\infty} f(z/y)\, g(y)\, dy / |y|$.

Expectations of Random Variables

$E(x) = \int_{-\infty}^{\infty} x\, f_x(x)\, dx$, provided that $\int_{-\infty}^{\infty} |x|\, f_x(x)\, dx < \infty$.

Example. Let $f_x(x) = 1 / \pi(1 + x^2)$. Then, E(X) does not exist.

Note. $E(X) = \int_0^{\infty} (1 - F(x))\, dx$ for a continuous variable, or $E(X) = \sum_{x=0}^{\infty} P(X \geq x)$.

Law of the Unconscious Statistician. Let $Y = g(X)$. Then, $E(Y) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$ or $E(Y) = \sum_{\text{all } x} g(x)\, P(X = x)$.

Note. Suppose $Y = a + b_1X_1 + b_2X_2 + \ldots$, where $E(X_i) = \mu_i$. Then, $E(Y) = a + b_1\mu_1 + b_2\mu_2 + \ldots$

Definition. The variance is given by $Var(X) = E((X - \mu_X)^2) = E(X^2) - E(X)^2$.

Defintion. The covariance is given by $Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - \mu_X\mu_Y$.

Note. If $X_1, \ldots, X_n$ are uncorrelated, then $Var(a + b_1X_1 + b_2X_2 + \ldots) = b_1\sigma_{X1}^2 + b_2\sigma_{X2}^2 + \ldots$ More generally, $Var(X + Y) = Var(X) + Var(Y) + 2\, Cov(X, Y)$.

Note. $\text{Cov}(X_1 + X_2, Y_1 + Y_2) = \text{Cov}(X_1, Y_1) + \text{Cov}(X_2, Y_1) + \text{Cov}(X_1, Y_2) + \text{Cov}(X_2, Y_2)$.
Note. $\text{Cov}(X, X) = \text{Var}(X)$.
Definition. The correlation coefficient is given by $\rho = \text{Cov}(X, Y) / \sqrt{\text{Var}(X)\,\text{Var}(Y)}$.
Note. $|\rho| \leq 1$. (Proof: Consider that $\text{Var}(X/\sigma_X + Y/\sigma_Y)$ must be positive.)
Chebyshev's Inequality. $P(|X - \mu_X| > t) < \sigma_X^2/t^2$.
Law of Large Numbers. Suppose X-bar $\sim [\mu_X, \sigma_X^2/n]$. Then X-bar converges to μ_x in probability.
Note. Since $P(|X - \mu_X| < 2\,\sigma_X) \geq 1 - \sigma_x^2/4\sigma_x^2$, at least 75% of observations will lie within two standard deviations of the mean.
Corollary. If $\text{Var}(X) = 0$, $P(X = \mu_X) = 1$.

Moment Generating Functions
Definition. If X is a random variable, then the moment-generating function is given by $M_X(t) = E(e^{tX})$.
Proposition. $M^{(r)}(0) = E(X^r)$
Proof. $M^{(r)}(t) = \int_{-\infty}^{\infty} x^r e^{tx} f_X(x)\,dx$. When t = 0, this is precisely $E(X^r)$.
Theorem. If M(t) exists in an open interval containing 0, then it uniquely determines f.
Proposition. Let Z = X + Y, with X and Y independent. Then, $M_Z(t) = M_X(t)M_Y(t)$.
Proposition. If Y = a + bX then $M_Y(t) = e^{at}M_X(t)$.
Example. If $X \sim N(\mu, \sigma^2)$, then $M_X(t) = e^{\mu t}e^{-(\sigma t)\wedge 2/2}$.
Central Limit Theorem. Let $X_1, X_2, \ldots$ be a sequence of random variables identically distributed with mean μ and variance σ^2 and moment generating function M defined in a neighborhood of 0. Then, $\lim_{n \to \infty} P((\sum X_i - n\mu)/\sigma\sqrt{n} \leq k) = \Phi(k)$.
Note. Because Gamma(α, λ), B(n, p), and Poisson(λ) can be considered sums, they are approximately normal for large values of α, n, and λ respectively.

Likelihood Functions
Definition. Suppose **X** is a random variable whose distribution depends on a parameter $\boldsymbol{\theta}$. Then, the likelihood function of θ given **x** is $L(\theta) = f_{\mathbf{X}}(\mathbf{x} \mid \theta)$. This measures how probable the data would be for a given value of θ.
Definition. The Bayes posterior density is $f_{\theta|\mathbf{X}}(\theta \mid \mathbf{x})$. In this, we think of the unknown θ as a random variable. If we call $p(\theta)$ the prior density, $f_{\theta|\mathbf{X}}(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta)p(\theta)/\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}|\theta)p(\theta)\,d\theta$.
Note. Notice that the denominator is a constant (with respect to θ). So the posterior density is proportional to $L(\theta)p(\theta)$. In particular, a flat prior (assuming a uniform distribution on possible values of θ) simply yields a posterior density of $L(\theta)$.

Large Sample Theory of Maximum Likelihood Estimators
Definition. $I(\theta_0) = E((\partial\,(\ln f(x|\theta_0)) / \partial\theta)^2) = -E(\partial^2\,(\ln f(x \mid \theta_0))/\partial\theta^2)$ is the information conveyed about the parameter of a distribution. θ_0 is the true parameter value and $x \sim f(x \mid \theta_0)$.
Theorem. The distribution of $\sqrt{[nI(\theta_0)]}\,(\theta^{\wedge} - \theta_0) \rightarrow N(0, 1)$, where θ_0 is the true value of the parameter and $\theta^{\wedge}$ is the estimate.

Definition. Suppose $X_i \sim f(X \mid \boldsymbol{\theta})$. Let $\mathbf{T} = \mathbf{T}(\mathbf{x})$. $\mathbf{T}$ is <u>sufficient</u> for $\boldsymbol{\theta}$ if $f_{\mathbf{X}}(\mathbf{x} \mid \mathbf{T})$ does not depend on θ.
Theorem. $\mathbf{T}$ is sufficient if and only if $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(\mathbf{T}, \boldsymbol{\theta})\, h(\mathbf{x})$ [the rest of the $\mathbf{x}$ can be factored out].
Note. If $\mathbf{T}$ is sufficient for $\boldsymbol{\theta}$, then the rest of the data is not needed to estimate.
Theorem (Rao-Blackwell). Let $\hat{\theta}$ be an unbiased estimator for θ. Then $\tilde{\theta} = E(\hat{\theta} \mid \mathbf{T})$ is an unbiased estimator with a smaller variance.
Definition. A sufficient statistic is <u>complete</u> if $E(g(\mathbf{T})) = 0$ if and only if g is identically 0.
Theorem. The lowest possible variance is achieved when $\mathbf{T}$ is complete. Then, this is called the unique minimum variance unbiased estimator (UMVU).
Theorem. An sufficient statistic from the exponential family is complete.
Theorem. Any two minimal sufficient statistics are functions of each other.

<u>Hypothesis Testing</u>
The Neyman-Pearson Testing Paradigm:

- Specify null and alternative hypotheses.
  - A hypothesis is simple if it specifies the values of all parameters
  - We define ω_0 to be the region of possible values of the parameters under the null hypothesis and ω_a to be the region under the alternative hypothesis.
- Specify a statistic of the data and the acceptance and rejection regions for the statistic.
- Type I Error: Rejecting the null hypothesis when it is true. The probability of type I error is controlled and called the <u>significance level</u>, α.
- Type II Error: Failing to reject the null hypothesis when it is false. The probability of *not* committing a type II error is called the <u>power</u> of the test. The Neyman-Pearson lemma states that, for two *simple* hypotheses, the likelihood ratio test is the most powerful. A test is <u>uniformly most powerful (UMP)</u> if it is most powerful in all cases covered by H_o and H_a.

Generalized Likelihood Ratio Test: Let $\Lambda = \max_{\theta \in \omega 0} \text{lik}(\theta) / \max_{\theta \in \Omega} \text{lik}(\theta)$, where H_0 specifies $\theta \in \omega_0$ and H_A specifies $\theta \in \Omega - \omega_0$. Reject for small values of Λ.

- The T-test and the χ^2-test are both GLRT's, since the p-values are small if and only if Λ is small.
- *Theorem.* Suppose $f(\mathbf{x} \mid \theta)$ is smooth as a function of θ and the mle of θ is consistent. Then, under H_0, the distribution of $-2 \log \Lambda$ approaches χ_k^2, where k = dim Ω - dim ω_0.

<u>Matrix Statistics</u>
Suppose $\mathbf{z} = \mathbf{c} + A\mathbf{y}$. Suppose $\mathbf{y} \sim [\boldsymbol{\mu}, \Sigma_{yy}]$. Then:

- $\Sigma_{zz} = A\Sigma_{yy}A^T$.
- $\boldsymbol{\mu}_z = \mathbf{c} + A\boldsymbol{\mu}_y$

Definition. U is a <u>pivotal quantity</u> (<u>pivot</u>) if $U = g(Y, \theta)$ has a density free of Y and θ.
Example. The t-statistic for data is a pivot.

Definition. If the density of Y is such that $f_{Y|\theta}(y-\theta|\theta)$ is free of y and θ, then θ is a pure location parameter and U = Y - θ is a pivot.

Example. If Y ~ N(μ, 1), then U = Y - μ ~ N(0, 1), and μ is a pure location parameter.

Definition. If the density of Y|θ is such that Y/θ is free of Y and θ, then θ is a pure scale parameter, and U = Y/θ is a pivot.

Example. Y | σ^2 ~ N(0, σ^2), then σ is a pure scale parameter.

Statistical Distributions

Order Statistics

Definition. X ~ Beta(α, β) if $f_X(x) = (\Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta))\, x^{\alpha-1}(1-x)^{\beta-1}$, $0 \le x \le 1$.

Note. $\int_0^1 x^{\alpha-1}x^{\beta-1}\,dx = \Gamma(\alpha)\Gamma(\beta) / \Gamma(\alpha+\beta)$.

Note. If X ~ Beta(α, β) then E(X) = α / (α + β).

Definition. Let $X_1, \ldots, X_n$ be identically distributed. Let $X_{(k)}$ be the k^{th} largest value (that is, $X_i < x_{(k)}$ for k – 1 X_i and $X_i > x_{(k)}$ for n – k X_i). Then, $X_{(k)}$ is called the k^{th} order statistic.

Fact. Suppose $X_1, \ldots, X_n$ ~ Uniform(0, 1). Then $X_{(k)}$ ~ Beta(k, n – k + 1). Hence, $E(X_{(k)}) = k / (n+1)$.

More generally:

$$f_{X(k)}(x) = P(X_i = x \text{ for some } i)P(X_i < x \text{ for } k-1 \text{ } X_i)P(X_i > x \text{ for } n-k \text{ } X_i)$$
$$= f_X(x)(F_X(x))^{k-1}(1 - F_X(x))^{n-k}\binom{n}{k-1,\,1,\,n-k}$$
$$= (\Gamma(n+1)/\Gamma(k)\Gamma(n-k+1))\, f_X(x)(F_X(x))^{k-1}(1 - F_X(x))^{n-k}$$

so the distributions may be plugged in to find the distribution of the order statistics.

Bivariate Normals

The bivariate normal distribution is given by $f_{XY} = (1/(2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)})\exp(-((X - \mu_X)^2/\sigma_X^2 + (Y - \mu_Y)^2/\sigma_Y^2 - 2\rho(X - \mu_X)(Y - \mu_Y)/\sigma_X\sigma_Y)/2(1 - \rho^2))$

- Then, Y | X ~ N($\mu_Y + \rho(\sigma_Y/\sigma_X)(X - \mu_X)$, $\sigma_Y^2(1 - \rho^2)$).
- $Cov(X, Y) = \rho\sigma_X\sigma_Y$
- We may consider generating U, V ~ N(0, 1) independently. With the proper choices of coefficients, we let X = aU + bV and Y = cU+dV, and these will generate any bivariate normal.

(The trick for normals: Factor the exponent into $(Y – A)^2/2V^2$; A is the mean and V is the variance. The leftover constants can be ignored.

Linear Predictions: Suppose we want to minimize $E((\alpha X + \beta - Y)^2)$ by predicting Y from X. We predict $Y^{\wedge} = \mu_Y + (\sigma_Y/\sigma_X)\rho(X - \mu_X)$.

Multivariate Normal Distributions

Definition. We say $\mathbf{X} = (X_1, \ldots, X_n)$ is multivariate normally distributed, $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, where Σ is an n × n matrix with non-negative eigenvalues, if $f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}\,|\Sigma|^{-1/2}\exp(-0.5\,(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}))$.

Note. If $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, the $\mathbf{Y} \sim N_m(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$

Definition. **X** is distributed as a standard multivariate normal if $\boldsymbol{\mu} = \mathbf{0}$ and Σ = I.

Note. Let $\Sigma^{-1/2}$ be defined by the inverse of $\Sigma^{1/2}$, where $\Sigma^{1/2}(\Sigma^{1/2})^T = \Sigma$. For any multivariate normal **X**, $\Sigma^{-1/2}\,(\mathbf{X} - \boldsymbol{\mu}) \sim N_n(\mathbf{0}, I)$. Note that $\Sigma^{1/2}$ and therefore $\Sigma^{-1/2}$ are not uniquely defined.

Example. If $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, then y_1 and y_2 are standard normal with correlation ρ.
Example. If Σ can be written in block form, the variables in one block are independent of the variables in the other block. Independent, identically-distributed variables have a covariance matrix of the form $\sigma^2 I$.
.

Statistics for Specific Situations

Testing for Goodness of Fit

Chi-Square Goodness of Fit Test: Suppose $X_1, \ldots, X_m$ are distributed according to a multinomial distribution. Then $\sum (X_i - E(X_i))^2 / E(X_i)$ approaches a χ_{m-1}^2 distribution.

- This may be used for grouped observations as well; combining categories improves the approximation, because the approximation is better with large $E(X_i)$.

Fitting Data Visually:

- Hanging Histogram: Draw a histogram of expected (n_j) – observed ($\hat{n_j}$).
  - Assuming $E(n_j - \hat{n_j}) = 0$, $Var(n_j - \hat{n_j}) \approx \hat{n_j}$.
- Hanging χ-gram: Plot $\chi = (n_j - \hat{n_j}) / \sqrt{\hat{n_j}}$
  - $Var(\chi) \approx 1$
- Hanging Rootgram: Plot $R = \sqrt{n_j} - \sqrt{\hat{n_j}}$
  - $Var(R) \approx Var(\sqrt{n_j}) \approx Var(n_j) / (2\sqrt{n_j})^2 = 1/4$

Testing for Normality

Quantile Plot: Take the inverse CDF's of the order statistics of the uniform (ie. $F^{-1}(k / (n+1))$); these are the "expected quantiles". Plot these against the ordered data – the line should be very straight.

Some other tests statistics: The distributions may be found through simulation.

- Skewness: $(1/n)(\sum(X_i - \text{X-bar})^3)/s^3$
- Kurtosis: $(1/n)(\sum(X_i - \text{X-bar})^4)/s^4$
- $\sum |\text{quantile} - E(\text{quantile})|$ or $\max \{|\text{quantile} - E(\text{quantile})|\}$

Survey sampling

Statification: Suppose we divide a population of size N into L strata based on some characteristic. Let N_l be the population size of the l^{th} stratum, $W_l = N_l/N$. Let n_l be the sample size from the l^{th} stratum.

- Estimating μ: $\text{X-bar}_s = \sum W_l \text{X-bar}_l = \sum_l W_l (\sum_i X_{il})/n_l$
- $Var(\text{X-bar}_s) = \sum W_l^2 (\sigma_l^2 / n_l) (1 - n_l/N_l)$
- The proportional allocation is given by $n_l = W_l n$. The variance of the estimate in this case is $\sum W_l \sigma_l^2 / n$.
- The Neyman allocation is given by $n_l = n W_l \sigma_l / \sum (W_k \sigma_k)$. Ignoring the finite population correction, this allocation minimizes variance, to be $(\sum W_l \sigma_l)^2/n$. This is $(\sum W_l (\mu_l - \mu)^2)/n$ less than the variance of the non-stratified estimate.
- Note that stratifying does not increase variance (assuming a good allocation). So stratifying on something irrelevant to the parameter being estimated is only useless. Ideally, strata have small variances within them and large variances between them.
- Stratifying based on sampling method can be useful for combining them properly.

Prior Information: Suppose μ_X is known and we want to know μ_Y. We may take a random sample and find X_i, Y_i for each element. We then estimate μ_Y as a function of μ_X, X-bar, and Y-bar.

- Difference Estimation: $\hat{\mu_Y}$ = Y-bar – k(X-bar - μ_X), where k is any constant.
  - $Var(\hat{\mu_Y}) = \sigma_Y^2/n + k^2\sigma_X^2/n – 2k\sigma_{XY}/n$
- Regression Estimation: Choose the k above by regressing y_i – y-bar on x_i – x-bar (no constant). This asymptotically minimizes the variance.
- Ratio Estimation: $\hat{\mu_Y}$ = Y-bar * μ_X / X-bar
  - This is biased: $E(\hat{\mu_Y}) - \mu_Y \approx (1/n)(1 – (n-1)/(N-1))(\sigma_X^2\mu_Y/\mu_X - \rho_{XY}\sigma_X\sigma_Y)/\mu_X$.
  - $Var(\hat{\mu_Y}) \approx (1/n)(1 – (n-1)/(N-1))((\mu_Y/\mu_X)^2\sigma_X^2 + \sigma_Y^2 – 2r\rho\sigma_x\sigma_y)$
    - $Var(\hat{\mu_y}) < Var(y\text{-bar}) \Leftrightarrow |\rho| > |r\ \sigma_x/\sigma_y|/2$, so this is useful when r is large relative to σ_x/σ_y.
- This can also be useful when the x_i are cheaper/easier to obtain (say, eye estimates). Then, a census of the x_i can be taken and a sample of the y_i (say, exact measurement) can be used to estimate μ_Y.

Non-response: Collect data from a sub-sample to non-respondents. Let y_1-bar be the mean for respondents and y_2-bar be the mean of the non-respondents. Then, we estimate y-bar as a weighted average of the respondents and the non-respondents (note that y_2 is weighted by the *total* number of non-respondents, not the number who were sampled to find y_2-bar). This increases the variance.

<u>Categorical Data Analysis</u>

One factor: Suppose there are k categories with π_i probability of an observation being the in the i^{th} category and y_i observations in each category.

- Mode: The category with the higheest y_i
- Concentration: a measure of spread
  - Genie Concentration: $V_G = \sum \pi_i(1 - \pi_i)$
  - Entropy: $V_E = \sum \pi_i \ln \pi_i$, where we assume $0 \ln 0 = 0$.
  - If all the observations are in one category, $V_E = V_G = 0$.
  - If the observations are evenly spread out: $\pi_i = 1/k$ for all i, and $V_G = (k-1)/k$ while $V_E = \ln k$.
    - If we normalize by dividing by the maximum value, V_E and V_G will be close.
- Measures of Association: Let $\pi_{ij} = P(x_1 \in i, x_2 \in j)$ and $\pi_{i*} = P(x \in i)$. x_1 and x_2 can have to do with different measurements in the same category. We assume j is known but i is not.
  - $\tau = (\sum\sum \pi_{ij}^2/ \pi_{i*} - \sum \pi_{*j}^2)/(1 - \sum \pi_{*j}^2)$
  - Uncertainty coefficient: $U = (\sum\sum \pi_{ij} \ln(\pi_{ij}/\pi_{i*}\pi_{*j}))/ \sum \pi_{*j} \ln(\pi_{*j})$
  - $\tau = U = 0$ if $\pi_{ij} = \pi_{i*}\pi_{*j}$ (ie. i and j are independent)
  - $\tau = U = 1$ if knowing j ensures that one knows i.
  - Cohen's kappa: $\kappa = (\sum \pi_{ii} - \sum \pi_{i*}\pi_{*i})/(1 - \sum \pi_{i*}\pi_{*i})$; measures the probability that two are in the same category.

Multinomial & Poisson Models: Suppose we have N things in k categories. Let X_n be the category of the n^{th} object and Y_i be the number of objects in the i^{th} category.

- $L(\boldsymbol{\pi} \mid \mathbf{x}) = \Pi_n \pi_{Xn} = \pi_1^{Y1}\pi_2^{Y2}\ldots\pi_k^{Yk}$
- $P(\mathbf{y} \mid N, k, \pi_1, \ldots, \pi_{k-1}) = n!\ \Pi\ \pi_i^{yi} / \Pi\ y_i!$
- $l(\pi_1, \ldots, \pi_{k-1}) = \textit{constant} + \sum y_i \ln \pi_i(\theta)$, where we may consider $\pi_i(\theta)$ to be a function describing the probabilities.
    - $l(\hat{\theta}) = \textit{constant} + \sum y_i \ln \hat{y_i}$ (where $\hat{y_i}$ is the expected value under $\hat{\theta}$)
- Poisson Model: Wait for a time t to observe k events.
    - P(event type i occurs in δT) = $\alpha_i \delta T$
    - Y_i = # of events of type i ~ Poisson($\alpha_i T$) = Poisson(μ_i)
    - Then, the number of events seen is a random variable.
    - $L(\boldsymbol{\mu} \mid \mathbf{x}) = \Pi\ P(X_n \text{ is of type i}) \propto e^{-\Sigma\mu i}\ \Pi\ \mu^{yi}/y_i!$
    - $l(\boldsymbol{\mu} \mid \mathbf{y}) = \textit{constant} - \sum \mu_i + \sum y_i \ln \mu_i$
    - $P(Y \mid \theta) = P(N = n)P(\mathbf{y} \mid N = n)$
        - This means that once we condition on the total number of observations, the Poisson model is still multinomial. Conversely, we may model multinomial data as though it came from a Poisson process.
- Sufficient Statistics: $\{y_1, \ldots, y_k\}$; N and k-1 of the y_i.
    - $\mu_i \mid n = E(y_i \mid n) = n\pi_i$, so $\mu_i = \pi_i\ E(n)$
- Non-Poisson multinomial: $P(\mathbf{y}) = P(N = n)P(\mathbf{y} \mid N = n)$, where P(N = n) may be any distribution. Note that the same results hold after we condition on n.

Two (and more) Way Tables

- Fisher's Exact Test
    - We assume that marginal probabilities (in both rows and columns) and count are fixed. In a 2×2 table, this allows exactly one parameter to vary (and affect the others). Fisher's exact test finds the distribution of this parameter under the constraints (which is hypergeometric, with some parameters), assuming that there is no interaction between the columns and rows. This allows the p-value to be found.
    - To use this on a larger table, combine rows or columns to create a 2×2 table. Or the test may be used, but the distributions must be found for multiple parameters.
- Pearson Goodness of Fit test
    - Expected Values: E_i = Count * Π Marginal Probabilities
    - Then, $\sum (O_i - E_i)^2/E_i \sim \chi^2$, with df = #cells - #parameters fixed

<u>ANOVA</u>

One-Way Layout: J_i measurements of a random variable Y_i are taken in each of I treatments ("levels").

- F-Test: Assume that $Y_i \sim N(\mu + \alpha_i, \sigma^2)$, with $\sum \alpha_i = 0$. Let Y_{ij} be the realizations of the Y_i.
    - $\sum\sum (Y_{ij} - Y_{..}\text{-bar})^2 = \sum\sum (Y_{ij} - Y_{i.}\text{-bar})^2 + \sum J_i (Y_{i.} - Y_{..}\text{-bar})^2$; we may write this as $SS_{TOT} = SS_W + SS_B$.
    - If $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$, then $SS_W/\sigma^2 \sim \chi^2_{I(\sum Ji\ -\ 1)}$. If $\alpha_i = 0$ for all i, $SS_B/\sigma^2 \sim \chi^2_{I-1}$ and is independent of SS_W.
    - If $\alpha_i = 0$ for all i, $F = (SS_B/(I-1))/(SS_W/(\sum J_i - I)) \sim F_{I-1,\ \sum Ji-I}$.

- Kruskal-Wallis Test: Rank all observations (averaging ranks for ties), replacing Y_{ij} by R_{ij}. We may use simulation, since we know all the values. Alternately, $K = 12\ SS_B / N(N+1) \sim \chi^2_{I-1}$. ($SS_B$ from the R_{ij}.)

Two-Way Layout: Suppose we classify measurements by two factors (A and B). Assume $Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$, where $\sum\alpha_i = \sum\beta_j = \sum\sum \delta_{ij} = 0$, and ε_{ijk} is independent of i, j. Suppose $Y_{ijk} \sim N(\mu + \alpha_i + \beta_j + \delta_{ij}, \sigma^2)$.

- $SS_{TOT} = \sum\sum\sum(Y_{ijk} - Y_{...})^2$; $SS_A = JK\sum_i (Y_{i..} - Y_{...})^2$; $SS_B = IK\sum_j (Y_{.j.} - Y_{...})^2$; $SS_{AB} = K\sum\sum (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2$; $SS_E = \sum\sum\sum (Y_{ij.} - Y_{ijk})^2$
- $SS_E / \sigma^2 \sim \chi^2_{IJ(K-1)}$. Under their respective null hypotheses, $SS_A/\sigma^2 \sim \chi^2_{I-1}$, $SS_B/\sigma^2 \sim \chi^2_{J-1}$, $SS_{AB}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$. This leads to more F tests.

The problem of multiple comparisons: If we want to test pairs for individual differences (instead of just finding an overall difference), we must use I(I-1)/2 tests. For a fixed α, the probability of making a type I error approaches 1 for large I.

- Tukey's Method: Use a distribution which adjusts for this.
- Bonferroni Method: Test all hypotheses at the $\alpha/(I(I-1)/2))$ level. This is conservative.

Randomized Block Design: Suppose there are J homogenous blocks. Assign each of the I treatments to one block. We may these assume there is no interaction. We model $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$. Our F-tests then use $F = (SS_A/(I-1))/(SS_{AB}/(I-1)(J-1))$ or $F = (SS_B/(J-1))/(SS_{AB}/(I-1)(J-1))$. (These statistics are conservative when there is interaction, since $SS_{AB}/(I-1)(J-1) > \sigma^2$ when there is interaction.)

Interaction Plot: Draw one line for each row category, connecting dots at each mean for each column category. If the lines are not parallel, there may be an interaction.

Random Effects Model: Suppose $Y_{ij} \mid è_i, ó_i^2 \sim N(è_i, ó_i^2)$, so that within group i, the measurements are similarly distributed but not like the measurements in other groups, so that $Y_{ij} = è_i + å_j$, and that $è_i \mid ì, A \sim N(ì, A)$ (or some random distribution).

- More generally: $è_i \mid \mathbf{â}, A \sim N(\mathbf{X}_i^T â, A)$ – this allows relationships among the $è_i$.
- If $\mathbf{Y}_{ij}$ are vectors, then $\mathbf{è}_i$ is also.
- Simplification: Assume $ó_i^2$ is known, so that all the Y_i.-bar are sufficient for $è_i$.
- Treat $ó_i^2 = V$ as constant over all i. Then, $Y_i \sim N(ì, V + A)$, so that $ì\ \hat{}$ is the average of all the Y_i and $A\hat{} = \max\{0, Ó(Y_i - Y\text{-bar})^2/k - V\}$ – if $A\hat{} = 0$, then $è_i$ is probably constant.

<u>Linear Regression</u>

Model: $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim [0, \sigma^2]$, independently of X.

- To estimate β_0 and β_1, we minimize $\sum (Y_i - \beta_0\hat{} - \beta_1\hat{} X_i)^2$.
- This yields unbiased estimators:
  - $\beta_0\hat{} = (\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i)/(n\sum X_i^2 - (\sum X_i)^2)$
  - $\beta_1\hat{} = (n\sum X_i Y_i - (\sum X_i)(\sum Y_i)) /(n\sum X_i^2 - (\sum X_i)^2)$
- Proving that they are unbiased includes the assumption that the X_i are fixed beforehand.
- Variances:
  - $\mathrm{Var}(\beta_0\hat{}) = \sigma^2\sum X_i^2/(n\sum X_i^2 - (\sum X_i)^2)$
  - $\mathrm{Var}(\beta_1\hat{}) = n\sigma^2/(n\sum X_i^2 - (\sum X_i)^2)$
  - $\mathrm{Cov}(\beta_0\hat{}, \beta_1\hat{}) = -\sigma^2\sum X_i/(n\sum X_i^2 - (\sum X_i)^2)$

    - To estimate σ^2: $s^2 = \sum(Y_i - Y_i^{\wedge})^2/(n-2)$ [more generally, n-k]
- This assumes homoskedasticity – if this is not true, let $Y' = f_1(Y)$ and $X' = f_2(X)$. If $f_1 = f_2$, then a straight line will stay straight, but the residuals will change, possibly to become homoskedastic.

Multiple Regression: Let $Y = (Y_1, \ldots, Y_n)^T$. Let $X = (x_{ij})$, where $x_{oj} = 1$ and x_{ij} is the j^{th} variable of the i^{th} observation. Let $\beta = (\beta_0, \ldots, \beta_{p-1})^T$ and $\varepsilon = (\varepsilon_1, \ldots \varepsilon_n)^T$. Then, $Y = X\beta + \varepsilon$.

- Then, $\beta^{\wedge} = (X^TX)^{-1}X^TY$. ($X^TX$ is invertible unless the x_i are perfectly multicollinear.) It is unbiased.
- $Var(\beta^{\wedge}) = \sigma^2(X^TX)^{-1}$.

Comparing Nested Models: Find the total sum of squares explained by the more restrictive model, the total explained by the less restrictive model, and the difference (error) between them. Find the degrees of freedom for each model (and the difference again). Let MS = SS/DF for restricted and error. Let F be the ratio of the MS's. Under the null hypothesis, $F \sim F_{treatment, error}$.

Residuals: Let $P = X(X^TX)^{-1}X^T$. Then, $Y^{\wedge} = PY$.

- $(I - P)^2 = (I - P)$
- $\sum(Y_i - Y^{\wedge}_i)^2 = ||(I - P)Y||^2 = Y^T(I - P)Y$
- $E(Y^T(I - P)Y) = \sigma^2(n - p)$
- An unbiased estimator for σ^2: $s^2 = \sum(Y_i - Y_i^{\wedge})^2/(n-p)$
- $\sum_{e^{\wedge}e^{\wedge}} = \sigma^2(I - P)$, so that the residuals are correlated.
    - We can standardize to $e_j^{\wedge}$/ sqrt((n-1)/n – $(x_j - x\text{-bar})^2/\sum(x_i - x\text{-bar})^2$))
- $E(Y^{\wedge}) = \beta_0 + \beta_1X$
- $Var(Y^{\wedge}) = \sigma^2(1/n + (x - x\text{-bar})/\sum(x_i - x\text{-bar})^2)$, so that the variance of the predictions gets larger as the x-value gets further from the mean. (This yields a <u>prediction interval</u>.)

R^2:

- R^2 = SSE / SST = $1 - \sum(y_i - y_i^{\wedge})^2/\sum(y_i - y\text{-bar})^2$. This is the proportion of total variation in y explained by the variation in x.
- In the one variable case, $r = \sum(x_i - x\text{-bar})(y_i - y\text{-bar})/\sqrt{\sum(x_i - x\text{-bar})^2\sum(y_i - y\text{-bar})^2} = \sqrt{R^2}$.
- As we add more predictors, the residual sum of squares will always decrease, so R^2 will always increase. So we adjust for this:
- $R\text{-bar}^2 = 1 - (n-1/n-p)(1 - R^2)$

Comparing two regressions (and two parts of data): $\{(x_i, y_i)\}$ and $\{(x_i', y_i')\}$.

- Let $s = \sqrt{(\sum(Y_i - Y_i^{\wedge})^2 + \sum(Y_i' - Y_i^{\wedge}')^2)/(n-2 + m-2)}$
- Let $t = (\beta_1 - \beta_1^{\wedge}) / s\sqrt{1/\sum(x_i - x\text{-bar})^2 + 1/\sum(x_i' - x\text{-bar}')^2}$
- Under the null hypothesis that the slpes are equal and the errors are normally distributed, $t \sim t_{n+m-4}$.
- Alternately, we may run a regression on all the data and test the residuals for patterns (t-tests, rank tests, run tests).

Interaction: Suppose Y depends on a variable X and an indicator (dummy) variable D. Then we may allow interaction if we use the regression $Y_i = \beta_0 + \beta_1X_i + \beta_2D_i + \beta_3X_iD_i$. This allows both the slope and the intercept to differ between the two groups. (If there are k categories, this requires k-1 dummy variables and more care insignificance tests.)

Inverting a Regression:

- Regress X on Y. (Assuming X is also random.)
- Calibration Interval: Solve $Y = \hat{\beta_0} + \hat{\beta_1} X$ for X. Then, the standard error is $s_{\hat{x}}$ = StandardError(Prediction of Y | $\hat{X}$)/$|\hat{\beta_1}|$.
- Find the prediction intervals for Y at each X and see which intervals contain the Y in question. (This is often wider than reversing a regression.)

X_i's are random: No change. (The variance of $\hat{\beta}$ changes, but not in a way that affects coverage probability of confidence intervals or anything like that.)

Measurement Error (in the X_i's): Now, $X_i \sim [\mu_X, \sigma_X^2]$. Estimates are unbiased, but variances of estimators are bigger.

Decision Theory and Bayesian Statistics

Terminology: An <u>action</u> is a choice made (from a set A). A <u>state of nature</u> is the true value of some unknown parameter. A <u>decision function</u>, d: X → A, maps an observation about the word to an action. A <u>loss function</u>, l(è, d(X)), measures the losses associated with taking a certain action in a certain state of the world. (Negative losses are gains…) The <u>risk function</u>, $R(è, d) = E_X(l(è, d(X)))$, measures the risk (expected loss) associated with a certain decision function.

Minimax Method: Minimize the maximum possible losses.

- Let $d^* = \min_{d \in D} (\max_{\theta \in \Theta} R(\theta, d))$, where D is the set of all possible decision functions and Θ is the set of all states of the world.

Bayes Rule: Assign a prior distribution to θ. Then, the Bayes risk is $B(d) = E_\theta(R(\theta, d))$. Minimize this.

- The posterior distribution is $h(\theta \mid X)$, so that the posterior Bayes risk is $E(l(\theta, d(X)) \mid X = x)$.
- To find a Bayes rule:
  - $h(\theta|X) = f(X|\theta)g(\theta)/(\int f(X|\theta)g(\theta)d\theta)$
  - $E(l(\Theta, a) \mid X=x) = \int l(\theta, a)h(\theta|X)d\theta$

0-1 Loss: The loss function is 0 if one is correct and 1 otherwise, so that $R(i, d) = 1 - P(d(X) = 1)$. This is the risk of being wrong.

Neyman-Pearson Lemma. Let d* be a test (decision rule) that accepts if $f(x|\theta_1)/f(x|\theta_2) > c$. Let α^* be the significance level of d*. Let d be any other test with significance level $\alpha \leq \alpha^*$. Then the power of d is at most the power of d*.

Proof. Let $c = (1 - \pi)/\pi$. Then, we accept when $\pi f(x|\theta_1)/(1-\pi)f(x|\theta_2) < 1$, so that this is actually a Bayes Rule with priors π, 1-π, and 0-1 loss.

Theorem. Suppose $R(\theta, d) = E((\theta - d(X))^2)$. Then the Bayes estimate, d(X), is the mean of the posterior distribution. However, if $R(\theta, d) = E(|\theta - d(X)|)$, then d(X) is the median. (In fact, all the M-estimates work this way.)

Definition. Let d_1, d_2 be two decision functions. d_1 <u>dominates</u> d_2 if $R(\theta, d_1) \leq R(\theta, d_2)$ for all θ. d_1 <u>strictly dominates</u> d_2 if d_1 dominates d_2 and the inequality is strict at any point. d_1 is <u>admissible</u> if it is not strictly dominated by an other decision function.

Theorem. Suppose d* is a Bayes Rule with respect to some prior, g, with $g(\theta) > 0$ for all θ and $R(\theta, d)$ continuous. The d* is admissible.

Proof. Use the fact that Bayes Rules minimize B(d).

Hierarchical Models: Suppose the observed variable depends on a parameter that varies itself (for each different observation) according to a different distribution. (This is akin to having a prior on the parameter.)

- Some distributions are conjugate, meaning that they fit well together.
  - Poisson-Gamma: $X_i \mid \theta_i \sim \text{Poisson}(\theta_i)$, $\theta_i \sim \text{Gamma}(\alpha, \lambda)$.
    - $X_i \mid \alpha, \lambda \sim \text{NegativeBinomial}(\alpha, \lambda/(1+\lambda))$.
    - $\theta_i \mid X_i, \alpha, \lambda \sim \text{Gamma}(X_i + \alpha, 1 + \lambda)$ – the expected value is a weighted average of the old parameters and the information derived from X_i's value; the weights depend on λ.
  - Beta-Binomial
  - Exposures and Covariates:

Bayesian Procedure (for Random Effects in ANOVA): Suppose there are k observations of Y_i, each with its own random effect.

- Specify priors on ì and A. (They may be non-informative – constant over all values.)
- Posterior densities:
  - If p(ì, A) is constant, then f(ì, A | Y) is proportional to f(Y | ì, A).
  - Let B = V/(V+A), so that $E(\grave{e}_i \mid Y_i, ì, A) = Bì + (1 - B)Y_i$. A flat prior for A is not flat for B. In fact, B | **Y** ~ ConstrainedGamma((k-3)/2, Ó$(Y_i - \text{Y-bar})^2$/2V, 1), which is a density if k > 3.

Stein Estimators: Suppose $X_i \mid \theta_i \sim N(\theta_i, V_i)$ and $\theta_i \sim N(\mu_i, A)$, for i = 1, …, k. Then $\theta_i \mid X_i \sim N((V_i\mu_i + AX_i)/(V_i + A), AV_i/(A + V_i))$, so that we estimate $\hat{\theta} = \mu_i + (1 - V_i/(A+V_i))(X_i - \mu_i)$. Let $d_S = (1 - c/\sum(X_i - \mu_i)^2)(X_i - \mu_i) + \mu_i$. If we assume a quadratic loss function, $R(\theta, d_S) = k + (-2c(k-2) + c^2)(E(1/\sum(X_i - \mu_i)^2))$. If $0 < c < k-2$, then the risk of this estimator is lower than of simply guessing $\theta_i = X_i$, even if the μ_i are badly chosen (though then $1/\sum(X_i - \mu_i)^2$ is very small). Best estimate: $\hat{\theta} = (1 - c/\sum(X_i - \mu_i)^2)(X_i - \mu_i) + \mu_i$. Choosing $\mu_i = X_i$ removes the degrees of freedom that were helping – that is bad.

Approximate Methods

Let $X \sim [\mu_X, \sigma_X^2]$. Let $Y = g(X)$. Then, using a Taylor expansion:

$Y = g(X) \approx g(\mu_X) + (X - \mu_X)\, g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''(\mu_X)$

$E(Y) \approx g(\mu_X)$ [this comes from the first term, or from the first two: $E(X - \mu_X) = 0$]

or, $E(Y) \approx g(\mu_X) + \frac{1}{2} E((X - \mu_X)^2) g''(\mu_X) = g(\mu_X) + \frac{1}{2}\sigma_x^2 g''(\mu_X)$

$\text{Var}(g(X)) \approx \sigma_X^2 (g'(\mu_X))^2$

These approximations are exact when g is linear or (for the mean) quadratic.

Simulation

<u>Generating Random Variables</u>

Inverse CDF Method: Suppose X is a random variable with CDF F_X. Then, the random variable $F_X(X)$ is distributed Uniform(0, 1). Inverting this, we find that if U ~ Uniform(0, 1) then $F_X^{-1}(U)$ has the same distribution as X.

Rejection Method: Suppose f_X is the density function of X. Let [a, b] be the (possibly infinite) interval on which $f_X(x) \neq 0$.

- Algorithm:
  - Choose a function M(x) such that $M(x) \geq f(x)$ on [a, b]. Let $m(x) = M(x) / \int_a^b M(x)\,dx$.
  - Generate T with density m(x).
  - Generate U ~ Uniform(0, 1), independent of T.
  - If $U * M(T) \leq f(T)$ then accept T. Otherwise, return to the first step.
- Proof of Method
  - $P(x \leq X \leq x + dx) = P(x \leq T \leq x + dx \mid \text{accept T}) = P(\text{accept T} \mid x \leq T \leq x + dx) * P(x \leq T \leq x + dx) / P(\text{accept T})$
  - $P(\text{accept T} \mid x \leq T \leq x + dx) = P(U \leq f(x) / M(x)) = f(x) / M(x)$
  - $P(x \leq T \leq x + dx) = m(x)$
  - $P(\text{accept T}) = \int_a^b (f(t)/M(t)) * m(t)\,dt = \int_a^b f(t) * (1 / \int_a^b M(x)\,dx)\,dt = 1 / \int_a^b M(x)\,dx$
  - Combining, we find $P(x \leq X \leq x + dx) = f(x)\,dx$, which is the correct distribution.
- We call the percentage of T's accepted the <u>acceptance rate</u>. This is exactly $\int_a^b f(x)\,dx / \int_a^b M(x)\,dx = 1 / \int_a^b M(x)\,dx$.

Another way to generate two normals:

- Let $X_1, X_2 \sim$ Uniform(-1, 1)
- Accept $x_1^2 + x_2^2 \leq 1$ (in the unit circle)
- Let $v = x_1^2 + x_2^2$
- $P(V \leq v) = P(X_1^2 + X_2^2 \leq v)$, so V ~ Uniform(0, 1)
- Let $R = \sqrt{(-2 \ln v)}$
- Let $\cos U = x_2 / \sqrt{v}$, $\sin U = x_1 / \sqrt{v}$.

Bootstrapping: A way to test without finding a distribution.

- Simulate the experiment (under the null hypothesis or with the given estimate) and compute the test statistic many times. Compare the actual test statistic to the ones found.
- Find a sufficient statistic and generate new data according to the distribution determined by it. This removes the dependence on the other parameters.
- Using simulation to find the distribution of parameters given the data: (Markov Chain Monte Carlo – Gibbs Sampling)
  - Let ö ~ (**è**, ì, A). Generate ö* by:
  - Drawing A* ~ A | Y, ì, **è**
  - Drawing ì* ~ ì | Y, A*, **è**
  - Drawing **è*** ~ **è** | ì*, Y, A*
  - Drawing ö* ~ (**è***, ì*, A*)
  - Note that ö* has the same distribution as ö, so that $ö_1, ö_2, ö_3, \ldots$ is a Markov chain.
    - *Definition.* $Y_0, Y_1, \ldots$ is a <u>first order Markov Chain</u> if $f(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Y_0) = f(Y_t \mid Y_{t-1})$
    - An <u>equilibrium distribution</u>, $f_0(Y)$ is a distribution that holds for all Y after a certain point in a Markov Chain. (If we can create a

Markov Chain with the equilibrium distribution, we may take random samples from that distribution.)

Gibbs Sampling: Let $\phi = (\phi_1, \ldots, \phi_p)$. Generate $\phi_i^{t+1} \sim f_1(\phi_i \mid Y, \phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_p)$ for each i in order (1 to p). These consecutive values are not independent, but they have the correct distribution. (Also, knowing when it is in equilibrium is hard.)

- Coupling: Start two processes at different initial points and update them with the same random numbers. When they agree, they had "forgotten" their starting points and reached equilibrium. (Can't start right after coupling, but close to then.)

EM Algorithm

Definition. Data with one variable of importance, y, are MAR (missing at random) if whether they are missing is independent of y. Data in which (x, y) are the variables of importance, x is always known, and y is missing are MAR if whether y is missing is independent of y conditional on x.

- Note that the probability that y is missing may depend on x and the data are still MAR.

The EM Algorithm:

- E-Step: Find the expectation of the sufficient statistics given the observed data and the current estimate of the parameter, $\theta^{(t)}$.
- M-Step: Find the maximum likelihood estimate of the parameter given the expectation of the sufficient statistics.
- The analysis:
  - $f(Y \mid \theta) = f(Y_{obs}, Y_{mis} \mid \theta) = f(Y_{mis} \mid Y_{obs}, \theta)\, f(Y_{obs} \mid \theta)$
  - $l(\theta \mid Y_{obs}) = l(\theta \mid Y) - \ln f(Y_{mis} \mid Y_{obs}, \theta)$, so $l(\theta \mid Y_{obs})$ does not depend on Y_{mis}.
  - Let $Q(\theta \mid \theta^{(t)}) = E_{Ymis}(l(\theta \mid Y_{mis}, Y_{obs})$
  - Let $H(\theta \mid \theta^{(t)}) = E_{Ymis}(\ln f(Y_{mis} \mid Y_{obs}, \theta)$
  - Then, $l(\theta \mid Y_{obs}) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)})$
  - *Theorem.* Suppose (1) $\partial\, (Q(\theta \mid \theta^{(t)}))/\partial\theta \mid_{\theta = \theta(t+1)} = 0$ for all t, (2) $\theta^{(t)}$ converges to θ^*, and (3) $f(Y_{mis} \mid Y_{obs}, \theta)$ is smooth as a function of θ. Then, $\partial(l(\theta \mid Y_{obs})/\partial\theta \mid_{\theta=\theta^*} = 0$, so that θ^* is a stationary point.
  - $I(\theta \mid Y_{obs}) = -\partial^2 Q(\theta_1 \mid \theta_2)/\partial\theta_1^2 + \partial^2 H(\theta_1 \mid \theta_2)/\partial\theta_1^2$; the observed information is the complete information (reflected in Q)minus the missing information (reflected in H).

Example: Suppose $Y \sim N(\mu, \sigma^2)$, $Y_{obs} = \{y_1, \ldots, y_m\}$ and $Y_{mis} = \{y_{m+1}, \ldots, y_n\}$. Recall that the sufficient statistics are $\sum y_i$ and $\sum y_i^2$.

- E Step:
  - $E(\sum_{i=1}^{n} y_i \mid \theta^{(t)}, Y_{obs}) = \sum_{i=1}^{m} y_i + (n\text{-}m)\, \mu^{(t)}$
  - $E(\sum_{i=1}^{n} y_i^2 \mid \theta^{(t)}, Y_{obs}) = \sum_{i=1}^{m} y_i^2 + (n - m)((\mu^{(t)})^2 + (\sigma^{(t)})^2)$
- M Step:
  - $\mu^{(t+1)} = E(\sum_{i=1}^{n} y_i \mid \theta^{(t)}, Y_{obs})/n = (\sum_{i=1}^{m} y_i + (n\text{-}m)\, \mu^{(t)})/n$
  - $(\sigma^{(t+1)})^2 = E(\sum_{i=1}^{n} y_i^2 \mid \theta^{(t)}, Y_{obs})/n - (\mu^{(t)})^2$
- In this case, we can solve for the limiting values (when $\mu^{(t)} = \mu^{(t+1)}$), to find that $\mu^{\wedge} = \sum_{i=1}^{m} y_i / m$ and $\sigma^{\wedge 2} = \sum_{i=1}^{m} y_i^2 / m - \mu^{\wedge 2}$.