Reference Sheet - Willy Chen (Latest Version)

Common Distributions

Normal $X \sim N(\mu, \sigma^2)$

$$PDF: \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$$
$$MGF: exp(\mu t + \frac{\sigma^2 t^2}{2})$$

Lognormal $X \sim Lognormal(\mu, \sigma^2)$

$$PDF: \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}\frac{(log(x)-\mu)^2}{\sigma^2}),\ x > 0$$
$$E[X] = exp(\mu + \frac{\sigma^2}{2}),\ Var(X) = [exp(\sigma^2) - 1]E[X]^2$$

Note: A lognormally distributed r.v. is an r.v. whose logged version is normally distributed.

Chi-Square $X \sim \chi^2_n$

Let $Z \sim Normal(0, I_n)$, $Z'Z = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$

$$E[X] = n,\ Var(X) = 2n$$

t Distribution with $df = n$

Let $Z \sim Normal(0,1)$, $X \sim \chi^2_n$. Define $T \equiv \frac{Z}{\sqrt{X/n}}$.
Then $T \sim \mathcal{T}_n$. As $\lim_{n\to\infty} \mathcal{T}_n \to Normal(0,1)$

F Distribution with $df = n$

Let $X_1 \sim \chi^2_{k_1}$, $X_2 \sim \chi^2_{k_2}$. Define $W \equiv \frac{X_1/k_1}{X_2/k_2} \sim \mathcal{F}_{k_1,k_2}$

Gamma $X \sim Gamma(\alpha, \beta)$

$$PDF: \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} exp(-\frac{x}{\beta}),\ x > 0$$
$$MGF: (1-\beta t)^{-\alpha},\ t < \frac{1}{\beta}$$
$$E[X] = \alpha\beta,\ Var(X) = \alpha\beta^2$$

When $\alpha = 1$, this is equivalent to $Exponential(\frac{1}{\beta})$.
If $X, Y \sim Gamma(\alpha_0, \beta_0)$, $X + Y \sim Gamma(2\alpha_0, \beta_0)$
$Gamma(\alpha = \frac{n}{2}, \beta = 2) \equiv \chi^2_n$.
α represents the time waiting and β represents the scale of the event (e.g. $\frac{1}{\beta}$ customers come in every α hours, $\lambda = \frac{\beta}{\alpha}$ for exponential).
Note: This distribution is typically used to model a continuous time until an event. However, generally, the **gamma distribution is NOT memoryless** unless it is the case of an exponential distribution. In a general question, try to use exponential instead (reducing α to 1. See problem $\star$ in selected problems for variations.

Exponential $X \sim Exponential(\lambda = \frac{1}{\theta})$

$$PDF: \lambda e^{-\lambda x},\ \lambda > 0\ = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$
$$CDF: 1 - e^{-\lambda x}\ = 1 - e^{-\frac{x}{\theta}}$$
$$MGF: \frac{\lambda}{\lambda - t},\ t < \lambda$$
$$E[X] = \frac{1}{\lambda},\ Var(X) = \frac{1}{\lambda^2}$$

Note: This distribution is typically used to model a continuous time until an event.
Exponential is memoryless[1]

Binomial $X \sim Binomial(n, p)$

$$PMF: \binom{n}{k} p^k (1-p)^{n-k}$$
$$MGF: (1 - p + pe^t)^n$$
$$E[X] = np,\ Var(X) = np(1-p)$$

Negative Binomial $X \sim NegBin(\mu, \alpha)$

$$\Gamma(r) = \int_0^\infty exp(-u) u^{r-1} du,\ r > 0$$
$$\Gamma(k) = (k-1)!,\ k \in \mathbb{Z}_{++}$$
$$PMF: \frac{\Gamma(\alpha + x)}{\Gamma(\alpha) x!} \left(\frac{\alpha}{\alpha+\mu}\right)^\alpha \left(\frac{\mu}{\alpha+\mu}\right)^x,\ x \in \mathbb{Z}_+$$
$$MGF: \left(1 + \frac{\mu}{\alpha}[1 - exp(t)]\right)^{-\alpha},\ t < -ln\left(\frac{\mu}{\alpha+\mu}\right)$$
$$E[X] = \mu,\ Var(X) = \mu + \frac{\mu^2}{\alpha}$$

When $\alpha = 1$, this is the *geometric* distribution
As $\alpha \to \infty$, NB converges to $Poisson(\mu)$

Poisson $X \sim Poisson(\theta)$

$$PMF: \frac{exp(-\theta)\theta^x}{x!},\ x \in \mathbb{N} \cup \{0\}$$
$$CDF: exp(-\theta) \sum_{x=0}^{t} \frac{\theta^x}{x!}$$
$$MGF: exp[\theta(exp(t) - 1)]$$
$$E[X] = \theta,\ Var(X) = \theta$$

$Poisson(\theta_1) + Poisson(\theta_2) = Poisson(\theta_1 + \theta_2)$
Note: This distribution is typically used to model the probability of an event happening given a specific time period. λ is the frequency of the event in said time period.
Poisson is memoryless[1].

Geometric $X \sim Geometric(p)$

k total trials ($k \in \mathbb{N}$)

$$PMF: (1-p)^{k-1} p$$
$$CDF: 1 - (1-p)^{\lfloor x \rfloor}$$
$$MGF: \frac{pe^t}{1 - (1-p)e^t},\ t < -ln(1-p)$$
$$E[X] = \frac{1}{p},\ Var(X) = \frac{1-p}{p^2}$$

k failures before success ($k \in \mathbb{N} \cup \{0\}$)
This is the sepcial case of $\Gamma(1, \mu)$

$$PMF: (1-p)^k p$$
$$CDF: 1 - (1-p)^{\lfloor k \rfloor + 1}$$
$$MGF: \frac{p}{1 - (1-p)e^t},\ t < -ln(1-p)$$
$$E[X] = \frac{1-p}{p},\ Var(X) = \frac{1-p}{p^2}$$

Geometric is memoryless[1]

Some Common Use Cases

Continuous wait time before an event:
$\Gamma(\alpha, \beta)$ or $Exponential(\lambda) \equiv \Gamma\left(1, \frac{1}{\lambda}\right)$
Discrete wait time before an event: $NegBin(\mu, \alpha)$ or $Geometric(\lambda)$
Probability of event in a given time:
$Poisson(\theta)$

Multivariate Normal Distribution

Conditional Normal

Consider random vectors $X_{m\times 1}, Y_{n\times 1}$ that are jointly normally distributed:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim Normal\left(\begin{pmatrix} \mu_X \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}\right)$$

where

$$\Sigma_{XY} = Cov(X, Y)_{m\times n} = \sum_{YX}^{'}$$

Then,

$$Y|X \sim Normal(\alpha + B'X, \Sigma_{Y|X})$$
$$B = \Sigma_{XX}^{-1}\Sigma_{XY}$$
$$\alpha = \mu_Y - B'\mu_X$$
$$\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

[1] For discrete $P(X > m + n \mid X \geq m) = P(X > n)$, for continuous $P(X > t + s \mid X > t) = P(X > s)$

Diagonalization of the Variance Matrix

A real, symmetric matrix Σ (which we assume variance matrices are), $\Sigma = QDQ'$ where Q is an orthonormal matrix ($QQ' = Q'Q = I$) and D is a diagonal matrix of eigenvalues. If we further assume that A is **positive definite**, then we can define $\Sigma^{-\frac{1}{2}} = QD^{-\frac{1}{2}}Q'$ where λ_i's are the eigenvalues and Q is made of corresponding eigenvectors.

$$D^{-\frac{1}{2}} = \begin{pmatrix} \lambda_1^{-\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \lambda_2^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & \lambda_n^{-\frac{1}{2}} \end{pmatrix}$$

If matrix B is symmetric and idempotent ($B^n = B$), then $X'BX = X'B'BX = (BX)'BX$.
If matrix $B_{n\times n}$ is symmetric, idempotent, and real with rank m ($\leq n$), it is diagonlizable with $B = QDQ'$ where D is a diagonal matrix with a total of m 1's in the diagonal.
$X \sum N(0, I_n) \Rightarrow X'_{1\times n} B_{n\times n} X_{n\times 1} \sim \chi^2_m$

Important Properties

$\sigma-$algebra

Let Ω be the outcome space and $\mathcal{B}$ be the $\sigma-$algebra generated by $\mathcal{B}$. Then $\mathcal{B}$ must satisfy:

1. $\Omega \in \mathcal{B}$
2. $\forall A \in \mathcal{B}$, $A^c \in \mathcal{B}$
3. $\forall i \in \mathbb{N}, A_i \in \mathcal{B}$, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$

Probability of Random Draws

	Without Replacement	With Replacement
Ordered	$P_k^n = \frac{n!}{(n-k)!}$	n^k
Unordered	$C_k^n = \frac{n!}{(n-k)!k!}$	$C_k^{n+k-1} = \frac{(n+k-1)!}{k!(n-1)!}$

Bayes' Rule

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
$$P(A|B)P(B) = P(B|A)P(A)$$
$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Probability as Expectation

Define the indicator function $I\{statement\}$ to be

$$I\{statement\} \equiv \begin{cases} 1 & \text{Statement is TRUE} \\ 0 & \text{Statement is False} \end{cases}$$

Then the probability of an event is the expectation of the indicator function of the event happening:

$$P(A) = E[I\{A\}]$$

Markov's Inequality

$$P(h(X) \geq b) \leq \frac{E[h(X)]}{b}$$

Chebyshev's Inequality

For $c > 0$, $a > 0$, $E[X^2] < \infty$

$$P(|X - \mu| \geq c) \leq \frac{\sigma_X^2}{c^2}$$
$$P(|X - \mu| \geq a\sigma) \leq \frac{1}{a^2}$$

Cauchy-Schwartz Inequality

$$|E[XY]| \leq E[|XY|] \leq [E[X^2]]^{\frac{1}{2}}[E[Y^2]]^{\frac{1}{2}}$$

Jensen's Inequality

Let $\mathcal{X} = supp(X)$, if $g : \mathcal{X} \to \mathbb{R}$ is **convex**, then

$$g(E[X]) \leq E[g(X)]$$

Interesting Property of Expectation

If $X \geq 0$, $E[X] = \int_{supp(X)} 1 - F(x)dx$

Law of Iterated Expectations

$$E_Y[Y] = E_X[E_Y[Y|X]] = E_X[E_Z[E_Y[Y|X, Z]|X]]$$

Law of Total Variance

$$Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$$

Conditional/Joint PDFs

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \iff X \perp\!\!\!\perp Y$$
$$f_X(x) = \int_{supp(Y)} f_{XY}(x, y)dy$$
$$f_{X|Y} = \frac{f_{XY}}{f_Y} = \frac{\int_{supp(Z)} f_{XYZ}dz}{f_Y}$$
$$= \int_{supp(Z)} \frac{f_{XYZ}(x, y, z)}{f_Y(y)} \cdot \frac{f_{XY}(x, y)}{f_{XY}(x, y)}dz$$
$$= \int_{supp(Z)} f_{Z|X,Y} \cdot f_{X|Y}dz$$

Moreover,

$$f_{Y,X|Z} = \frac{f_{YXZ}(y, x, z)}{f_Z(z)} = \frac{f_{Y|X,Z}(y|x, z)f_{X,Z}(x, z)}{f_Z(z)}$$
$$= f_{Y|X,Z}(y|x, z) \cdot \frac{f_{X,Z}(x, z)}{f_Z(z)}$$
$$= f_{Y|X,Z}(y|x, z)f_{X|Z}(x|z)$$

Matrix Algebra

A $n \times n$ matrix A is orthogonal if $A^TA = I_n$.
A $n \times n$ matrix A is idempotent if $\forall n \in \mathbb{N}$, $A^n = A$.
Two matrices A_1, A_2 are orthogonal to each other if $A_1A_2 = 0_n$
If a matrix $Q_{n\times k}$ is idempotent, then $Rank(Q) = tr(Q)$
For any two matrices $A_{n\times k}, B_{k\times l}$, we have $tr(AB) = tr(BA)$, $tr(A + B^T) = tr(A) + tr(B), tr(cA) = c \cdot tr(A), c \in \mathbb{R}$

$$X^\top \iota\iota^\top X = (\sum_{i=}^{n} X_i)^2$$
$$\iota^\top \iota = n$$
$$\bar{X} = (\iota^\top \iota)^{-1}\iota^\top X$$
$$\vec{\bar{X}} = \iota(\iota^\top \iota)^{-1}\iota^\top X$$
$$P_X = X(X^\top X)^{-1}X^\top$$
$$M_X = I - P_X$$

Trace can be rearranged cyclically

$$tr(ABCD) = tr(DABC) = tr(CDAB) = tr(BCDA)$$

Asymptotic Properties

- A sequence of random variables X_n converges **in mean squared errors** to a random variable X if $E[(X_n - X)^2] \to 0$
- A sequence of random variables X_n converges **in probability** to a random variable X if $\forall \varepsilon > 0$, $P(|X_n - X| > \varepsilon) \to 0$. We can also denote this as $X_n \xrightarrow{p} X$ or say $X_n - X$ is $o_p(1)$.
- A sequence of random variables X_n converges **in distribution** to a random variable X if $F_{X_n} \to F_X$.
- A sequence of random variables X_n is **bounded in probability** if $\forall \varepsilon > 0$, $\exists b_\varepsilon > 0$, $P(|X_n| \geq b_\varepsilon) \leq \varepsilon$. We denote X_n being bounded in probability as $X_n = O_p(1)$.
- Note that convergence in probability implies convergence in distribution, which implies boundedness in probability.
- Asymptotic Unbiasedness + Vanishing Variance $\Rightarrow$ Consistency
- Consistency + Bounded Variance $\Rightarrow$ Asymptotic Unbiasedness

Weak Law of Large Numbers (WLLN)

Version 1: If $X_i \sim D_i(X)$ such that $E[X], E[X^2] < \infty$ AND $E[X_iX_j] = 0$, then $\bar{X}_n \xrightarrow{p} E[X]$

Version 2: If $X_i \overset{iid}{\sim} D(X)$ such that $E[X] < \infty$, then $\bar{X}_n \xrightarrow{p} E[X]$

Central Limit Theorem (CLT)

If $X_i \overset{iid}{\sim} D(X)$ such that $E[X_i^2] < \infty$, then

$$\sqrt{n}(\bar{X_n} - \mu) \xrightarrow{d} N\left(0, Var(X_i)\right)$$

Continuous Mapping Theorem (CMT)

Let $g(X) : supp(X_n) \to supp(X)$. If g is a continuous function on the support of X, we have

$$X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$$
$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$
$$X_n \xrightarrow{p} c \in \mathbb{R} \wedge Y_n \xrightarrow{p} Y \Rightarrow X_n Y_n \xrightarrow{p} cY$$
$$X_n \xrightarrow{p} c \in \mathbb{R} \wedge Y_n \xrightarrow{d} Y \Rightarrow X_n Y_n \xrightarrow{d} cY$$

Stochasic Order Algebra

(i) $o_p(1) + o_p(1) = o_p(1)$

(ii) $o_p(1) + O_p(1) = O_p(1)$

(iii) $o_p(1) \cdot O_p(1) = o_p(1)$

(iv) $(1 + o_p(1))^{-1} = O_p(1)$

(v) $o_p(O_p(1)) = o_p(1)$

Properties of Estimators

- An estimator $\hat{\theta}$ is biased if $E[\hat{\theta}] \neq \theta$.
- An estimator $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{p} \theta$
- An estimator is asymptotically unbiased if $Bias(\hat{\theta}) \to 0$
- In general, $E[\bar{X}_n] = E[X]$, $Var(\bar{X}_n) = \frac{Var(X)}{n}$
- The best unbiased estimator is an estimator that is unbiased AND has the smallest asymptotic variance
- The best unbiased estimator is an estimator that is unbiased AND has the smallest asymptotic variance, and is a linear function of the observations X_i.
- A sequence of estimators $\hat{\theta}_n$ is $\sqrt{n}-$**consistent** if

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_p(1)$$

- Two sequences of estimators $\hat{\theta}_n, \tilde{\theta}_n$ is $\sqrt{n}-$**asymptotically equivalent** if

$$\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) = O_p(1)$$

The Delta Method

For a sequence of estimator $\hat{\theta}_n$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, v(\theta)\right)$$

and that $g(x)$ is continuous on $supp(\theta_n)$, we have

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} N\left(0, g'(\theta) \cdot v(\theta) \cdot g'(\theta)^\top\right)$$

Cookbook Approach to Delta Method

Suppose that we want to find the asymptotic distribution of an estimator $\hat{\theta}$:

Step 1: $\hat{\theta} = \bar{X}$?

Yes $\Rightarrow$ CLT, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, Var(X))$

No $\Rightarrow$ Step 2

Step 2: Is $\hat{\theta}$ a function of $\bar{X}_n$?

Yes $\Rightarrow$ CLT, $\sqrt{n}(\bar{X} - E[X]) \xrightarrow{d} N(0, Var(X))$

and by Delta method

$\sqrt{n}(g(\bar{X}) - g(E[X])) \xrightarrow{d} N(0, \nabla g^\top(\theta) Var(X) \nabla g(\theta))$

No $\Rightarrow$ Step 3

Step 3: Is $\hat{\theta}$ a function of some $\bar{Y}_n$? Most likely yes, $\Rightarrow$ CLT,

$\sqrt{n}(\bar{Y} - E[Y]) \xrightarrow{d} N(0, Var(Y))$

and by Delta method

$\sqrt{n}(g(\bar{Y}) - g(E[Y])) \xrightarrow{d} N(0, \nabla g^\top(\theta) Var(Y) \nabla g(\theta))$

No, then we likely cannot use the delta method.

Common Estimators

Method of Moments: Figure out which moment you want to estimate, then use the sample analogue as the estimator. (See Example).

Maximum Likelihood Estimators: Figure out the joint (log-)likelihood function of the $n-$sample, check first and second order conditions so that you have an estimator that maximizes the joint likelihood function. (Note that MLE are usually consistent asymptotically most efficient, but they are often biased.)

Common MLE for First moments:

Poisson	$\bar{X}_n$
Exponential	$\bar{X}_n$
Normal	$\bar{X}_n$
Neg. Bin./Geometric	$\bar{X}_n$
Binomial/Bernoulli	$\bar{X}_{nm}/\bar{X}_n$
Uniform	max $\lvert X_i \rvert$ or max X_i

Common Estimators for Variance:

Normal	$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum(X_i - \bar{X})^2$
General	$\hat{\sigma}^2_{Unbiased} = S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$

Cramer-Rao Lower Bound

For any distribution that satisfies:

1. $f(x;\theta)$ has bounded support in x and the bounds do not depend on θ (so CRLB does not work on uniform)
2. $f(x;\theta)$ has infinite support, is continuously differentiable, and integrable for all θ

The lower bound of the asymptotic variance of **any estimator** for parameters of the distribution is

$$V(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is the Fisher information matrix defined as:

$$I(\theta) = nE\left[\left(\frac{\partial l(X;\theta)}{\partial \theta}\right)^2\right] = -nE\left[\frac{\partial^2 l(X;\theta)}{\partial \theta^2}\right]$$

where $l(X;\theta)$ is the log-likelihood function for a single observation.

Notice that, by construction, $\hat{\theta}_{MLE}$ always achieves CRLB, but it is also almost always biased.

Pooled Standard Deviation Estimator

Suppose X and Y are assumed to have the same variance and $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$, and we have the null hypothesis $H_0 : \mu_x = a\mu_y$ against $H_1 : \mu_X \neq \mu_y$.

Suppose that σ^2 is known, then for asymptotic inference, we can use the standard normal distribution. By CLT, we have:

$$\sqrt{n_X}(\bar{X}_{n_X} - \mu_X) \xrightarrow{d} N(0, \sigma^2)$$
$$\sqrt{n_Y}(\bar{Y}_{n_Y} - \mu_Y) \xrightarrow{d} N(0, \sigma^2)$$

meaning

$$\bar{X}_{n_X} \overset{a}{\sim} N(\mu_X, n_X^{-1}\sigma^2)$$
$$\bar{Y}_{n_Y} \overset{a}{\sim} N(\mu_Y, n_Y^{-1}\sigma^2)$$

Under H_0, we have

$$\bar{X}_{n_X} - a\bar{Y}_{n_Y} \overset{a}{\sim} N(0, (n_X^{-1} + a^2 n_Y^{-1})\sigma^2)$$

so

$$\frac{\bar{X}_{n_X} - a\bar{Y}_{n_Y}}{\sigma\sqrt{n_X^{-1} + a^2 n_Y^{-1}}} \overset{a}{\sim} N(0,1) \ (\star)$$

Now suppose that σ^2 is unknown, meaning that we would have to estimate it somehow. Naturally, since we now have to estimate 2 parameters from the normal distribution, we want to construct a test-statistic T that follows a $T-$distribution:

$$\frac{N(0,1)}{\sqrt{\chi^2/n}} = T \sim T_n$$

Notice that the estimator $(\star)$ above can conveniently serve as the numerator of our test statistic. Next, we have to estimate the standard deviation of the two population. Since we assumed that the variances across the two distributions are equal, we can use a pooled estimator.

Recall that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ is the unbiased population estimator because the analogue sample variance estimator is biased:

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{n-1}{n}\sigma^2$$

So we have

$$E\left[\frac{1}{n_X}\sum_{i=1}^{n_X}\left(X_i - \bar{X}\right)^2\right] = \frac{n_X - 1}{n_X}\sigma^2$$

This means that our new estimator S^2 has to be the weighted average of the two estimators $S^2 = w_X \sum_{i=1}^{n_X} \left(X_i - \bar{X}\right)^2 + w_Y \sum_{i=1}^{n_Y} \left(Y_i - \bar{Y}\right)^2$. Consider what would happen if I multiplied each estimator by their weights (i.e., n_X and n_Y)

$$E\left[\sum_{i=1}^{n_X} \left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{n_Y} \left(Y_i - \bar{Y}\right)^2\right] = (n_X - 1)\sigma^2 (n_Y - 1)\sigma^2 = (n_X + n_Y - 2)\sigma^2$$

and so we obtain a pooled unbiased population variance estimator:

$$\frac{1}{n_X + n_Y - 2} E\left[(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2\right] = \sigma^2$$

$$\Rightarrow E[S^2] = E\left[\frac{n_X - 1}{n_X + n_Y - 2} S_X^2 + \frac{n_Y - 1}{n_X + n_Y - 2} S_Y^2\right] = \sigma^2$$

Now we need to determine the asymptotic variances of S^2, recall that, by CLT and some algebra, we can show that

$$\frac{X_i - \bar{X}_{n_X}}{\sigma} \overset{a}{\sim} N(0,1)$$

and we can rewrite S^2 (approximately) as:

$$S^2 = (n_X + n_Y - 2)^{-1} \cdot \begin{pmatrix} X_1 - \bar{X} \\ \cdots \\ X_{n_X} - \bar{X} \\ Y_1 - \bar{Y} \\ \cdots \\ Y_{n_Y} - \bar{Y} \end{pmatrix}^T \underbrace{\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & I_{n_X+n_Y-2} & \vdots \\ 0 & \cdots & 0 \end{pmatrix}}_{Rank = n_X + n_Y - 2} \begin{pmatrix} X_1 - \bar{X} \\ \cdots \\ X_{n_X} - \bar{X} \\ Y_1 - \bar{Y} \\ \cdots \\ Y_{n_Y} - \bar{Y} \end{pmatrix}$$

So we know that $(n_X + n_Y - 2)\frac{S^2}{\sigma^2} \sim \chi^2_{n_X+n_Y-2}$, and hence our $T-$ statistic is

$$\frac{\star}{\sqrt{((n_X + n_Y - 2)S^2/\sigma^2)/(n_X + n_Y - 2)}} = \frac{\bar{X}_{n_X} - a\bar{Y}_{n_Y}}{\sqrt{S^2}\sqrt{n_X^{-1} + a^2 n_Y^{-1}}} = t \sim T_{n_X+n_Y-2}$$

Hypothesis Testing

When we observe data, we create a test statistics T_n. Putting T_n against a critical region C_α given the pre-determined *size* of the test α. We reject the null hypothesis if $T_n \in C_\alpha$

- **Size:** $\alpha = P(Reject \mid H_0) = P(T_n \in C_n \mid H_0$
- **Power:** $1 - \beta = \beta(\theta) = P(Reject \mid H_0)$

For a test, we want to maximize power ($1 - \beta$ or equivalently $\beta(\theta)$) given a specific α.

p-value is the minimal α needed to reject H_0 with the data observed. One should not think about this as a probability. Formally,

$$p - value \equiv \inf\{\alpha \in (0,1) \mid T_n \in C_\alpha\}$$

This definition is important because it gives clear guidelines on calculating p-values for a discrete R.V.
A sequence of tests is **consistent** if it has asymptotic power $\lim_{n\to\infty} \beta(\theta) = \lim_{n\to\infty} P(reject\, H_0 \mid \theta) = 1, \forall \theta \in \Theta_1$

Local Power Analysis

Consider, instead, a sequence of alternative hypotheses $H_{n1} : \mu_n = \mu_0 + \frac{\delta}{\sqrt{n}}$ against the $H_0 : \mu = \mu_0$. Then we can calculate the power of a sequence of tests $\beta(\theta) = P(reject\, H_0 \mid H_{n1})$. As $n \to \infty$, H_1 gets closer and closer to H_0. This gives us an idea of what the local power looks like when designing a test, **not designing an estimator**.

Confidence Intervals

By CLT, if we have iid samples with finite second moment, we know that we can achieve some type of asymptotic normality of our estimators. For example, say

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, v(\theta))$$

then we can also write

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\sqrt{v(\theta)}}\right) \xrightarrow{d} N(0,1)$$

This means that we can use the $z-$table to form the interval F_α for evidence that we would **fail to reject under** H_0 as:

$$F_\alpha = (\mu_0 - z_\alpha se(\hat{\theta}))$$

such that $T_n \in F_\alpha \Rightarrow$ Fail to reject H_0.
Example:
For $H_0 : \beta_k = a$ against $H_1 : \beta_k < a$, the confidence interval is

$$\left(-\infty, \hat{\beta}_k + c_\alpha \cdot se(\hat{\beta}_k)\right]$$

Common Formulas

Given the regression model

$$y = X\beta + u = \beta_1 X_1 + \beta_2 X_2 + u$$

where ι is contained in X_1, we can write

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$
$$\hat{\beta}_1 = (X_1^\top M_2 X_1)^{-1} X_1^\top M_2 y$$
$$\hat{u} = M_X y$$

Let the general linear estimator of β be denoted $\beta^* = (W^\top X)^{-1} W^\top y$ where W is non-random, then

$$E[\beta^* | X] = \beta + (W^\top X)^{-1} W^\top E[u|X]$$

$$\begin{aligned} Var(\beta^* | X) &= Var((W^\top X)^{-1} W^\top u | X) \\ &= (W^\top X)^{-1} W^\top Var(u|X) W (X^\top W)^{-1} \\ Var(\beta_2 | X) &= \frac{\sigma_0^2}{TSS_1(1 - R_1^2)} \\ Var(R\hat{\beta}|X) &= RVar(\hat{\beta}|X)R^\top = \sigma_0^2 R(X^\top X)^{-1} R^\top \\ se(\hat{\beta}|X) &= \sqrt{\frac{1}{n-k} y^\top M_X y (X^\top X)^{-1}} \\ &= \sqrt{\frac{1}{n-k} \hat{u}^\top \hat{u} (X^\top X)^{-1}} \end{aligned}$$

When testing 1 linear restriction (under GM1-5), we use

$$t = \frac{R\hat{\beta} - r}{se(\hat{Rbeta}|X)} = \frac{R(X^\top X)^{-1} X^\top u}{\sqrt{\frac{1}{n-k} \hat{u}^\top \hat{u} R(X^\top X)^{-1} R^\top}}$$
$$= \frac{1}{\sqrt{\frac{1}{n-k} \underbrace{\frac{u^\top}{\sigma_0} M_X \frac{u}{\sigma_0}}_{\sim \chi^2_{n-k}}}} \cdot \frac{R(X^\top X)^{-1} X^\top u}{\underbrace{\sqrt{\sigma_0^2 R(X^\top X)^{-1} R^\top}}_{\sim N(0,1)}} \sim T_{n-k}$$

When testing r linear restriction (under GM1-5), we use

$$W = (R\hat{\beta} - r)^\top \left[\hat{Var}(R\hat{\beta} - r)\right]^{-1} (R\hat{\beta} - r)$$
$$= (R\hat{\beta} - r)^\top \left[s^2 R(X^\top X)^{-1} R^\top\right]^{-1} (R\hat{\beta} - r)$$
$$= \frac{(R\hat{\beta} - r)^\top \left[\sigma_0^2 R(X^\top X)^{-1} R^\top\right]^{-1} (R\hat{\beta} - r)}{s^2/\sigma_0^2} \sim (n-k)\frac{\chi_r^2}{\chi_{n-k}^2}$$

$$F = \frac{1}{r} W \sim F_{r,n-k}$$

Gauss-Markov Assumptions

(GM.1) Specification is correct

(GM.2) X has full rank

(GM.3) $E[u|X] = 0$

(GM.4) $Var(u|X) = E[uu^\top | X] = \sigma_0^2 I_n$

(GM.5) $u|X$ is distributed normally

With the first 4, $\hat{\beta}$ is BLUE. With 5, we can get the exact distribution of the t statistic and OLS is also an MLE.

OLS Asymptotics Assumptions

(OLS.1) Specification is correct

(OLS.2) The sequences $(Y_t), (X_t)$ are sampled i.i.d. or has covariance weak enough for CLT to hold

(OLS.3) $\frac{1}{n} \cdot X^\top X \to E[X_t^\top X_t]$ which has full rank

(OLS.4) $E[u_t | X_t] = 0$

(OLS.5) $E[u_t^2 X_t^\top X_t]$ has full rank