Advanced Statistics
Spring 2022

Probability and Linear Algebra Review

Dr. Alon Kipnis
Material credit: Art Owen

Probability

Probability and Random Variables

- Probability space $(\Omega, \mathcal{F}, \Pr)$
  - $\mathcal{F}$ is a σ-field if:
    - $\Omega \in \mathcal{F}$
    - $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$
    - $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$
  - $\Pr : \mathcal{F} \to [0, 1]$ is a probability measure if:
    - $\Pr(A) \geq 0$, $A \in \mathcal{F}$
    - $\Pr(\Omega) = 1$
    - $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint $\Rightarrow \Pr(\cup_i A_i) = \sum_i \Pr(A_i)$
- Random variable: function $X \ : \ \Omega \to \mathbb{R}$ such that $\{\omega \ : \ X(\omega) \leq a\} \in \mathcal{F}$
- Random vector: function $X \ : \ \Omega \to \mathbb{R}^n$ such that $\{\omega \ : \ X_i(\omega) \leq a\} \in \mathcal{F}$, $i = 1, \ldots, n$
- Notation

$$X \leq a := \{\omega \ : X(\omega) \leq a\},$$

so that

$$\Pr(X \leq a) = \Pr\left(\{\omega \ : X(\omega) \leq a\}\right)$$

Independence and Bayes' Law

- **Events** $A, B \in \mathcal{F}$ are **independent** iff
  $\Pr(A, B) = \Pr(A \cap B) = \Pr(A) \Pr(B)$
- **Random variables** X and Y are **independent** iff
  $\Pr(X \leq a, Y \leq b) = \Pr(X \leq a) \Pr(Y \leq b)$ for any $a, b \in \mathbb{R}$
- The **conditioned probability** of $A \in \mathcal{F}$ given $B \in \mathcal{F}$ is

$$\Pr(A|B) \frac{\Pr(A \cap B)}{\Pr(B)}, \qquad A \in \mathcal{F}$$

- Bayes' law:

$$\Pr(A|B) = \frac{Pr(B|A)P(A)}{P(B)}$$

Distribution Functions

- The **cumulative distribution function** (CDF) of the RV X:

$$F_X(x) := \Pr[X \leq x] = \Pr[\{\omega, : X(\omega) \leq x\}], \qquad x \in \mathbb{R}$$

- The **probability density function** (PDF) of the RV X, if exists, satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

- The **multivariate CDF** of the d-dimensional random vector X is the function $F_X : \mathbb{R}^d \to [0, 1]$

$$F_X(x_1, \dots, x_d) := \Pr\left[X_1 \leq x_1, \dots, X_d \leq x_d\right]$$

- The **multivariate PDF** of the d-dimensional random vector X, if exists, is the function $f_X : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$F_X(x_1, \dots, x_d) := \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_X(t_1, \dots, t_d)dt_1 \cdots dt_d,$$

$(x_1, \dots, x_d) \in \mathbb{R}^d$.

- The **quantile** function of the RV X is

$$Q(p) := \inf\left\{x \in \mathbb{R} \,:\, p \leq F(x)\right\}, \qquad p \in [0, 1]$$

Expectation and Moments

Suppose that the RV X has a density function f_x, and let $h(x)$ be a real valued function on $\mathbb{R}$.

- The expectation of $h(X)$ is

$$\mathbb{E}\left[h(X)\right] := \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

  provided the integral exists. Otherwise, $\mathbb{E}\left[h(X)\right]$ does not exists.
- Taking $h(x) = x^k$ gives the k-th moment of X
- Some special moments of interest have given names:
  - The **mean** $\mu = \mathbb{E}\left[X\right]$ corresponds to $h(x) = x$
  - The **variance** of X is $\sigma^2 := \mathbb{E}\left[(X-\mu)^2\right]$
  - The **skewness** of X is $\gamma := \mathbb{E}\left[(X-\mu)^3\right]/\sigma^3$
  - The (excess) **kurtosis** of X is $\kappa := \mathbb{E}\left[(X-\mu)^4\right]/\sigma^4 - 3$
- γ is useful as a measure of symmetry; it is zero for symmetric distributions
- $\kappa = 0$ when $X \sim N(0,1)$. κ is useful in measuring whether the tails of the distribution are heavier ($\kappa > 0$) or lighter ($\kappa < 0$) than the tails of the normal distribution.

Expectation and Moments

These moments behave nicely under averages. Suppose that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ for X_i iid. Then

- $\mu(\bar{X}) = \mu$
- $\sigma^2(\bar{X}) = \sigma^2/n$
- $\gamma(\bar{X}) = \gamma/\sqrt{n}$
- $\kappa(\bar{X}) = \kappa/n$

From the CLT we expect $\gamma(\bar{X}) \to 0$ and $\kappa(\bar{X}) \to 0$. The evaluation above shows that the heaviness of the tail, as measured by κ, approaches that of the normal distribution much quicker than the skewness. For this reason, we expect the normal approximation resulted from the CLT to apply more accurately for symmetric distributions. When dealing with non-normality, skewness is more of an issue than non-Gaussian tails.

Random Vectors and Matrices

- Let X be an $n \times p$ matrix of random variables

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

- The **expectation** of X is defined as

$$\mathbb{E}[X] := \begin{pmatrix} \mathbb{E}[X_{11}] & \mathbb{E}[X_{12}] & \cdots & \mathbb{E}[X_{1p}] \\ \mathbb{E}[X_{21}] & \mathbb{E}[X_{22}] & \cdots & \mathbb{E}[X_{2p}] \\ \vdots & & \ddots & \vdots \\ \mathbb{E}[X_{n1}] & \mathbb{E}[X_{n2}] & \cdots & \mathbb{E}[X_{np}] \end{pmatrix}$$

- Taking $p = 1$ or $n = 1$, gives the definition for the expected value of row or columns vectors, respectively.
- Note that, for non-random $A \in \mathbb{R}^{* \times n}$ and $B \in \mathbb{R}^{p \times *}$,

$$\mathbb{E}[AX] = A\mathbb{E}[X], \qquad \mathbb{E}[XB] = \mathbb{E}[X]B$$

Covariances

Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be two random column vectors.

- The **covariance of X and Y** is

$$\mathrm{Cov}(X, Y) := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top\right] \in \mathbb{R}^{n \times m}$$

  (the i, j-th coordinate of $\mathrm{Cov}(X, Y)$ equals $\mathrm{Cov}(X_i, Y_j)$)
- The **variance-covariance matrix** of X is

$$\mathrm{Var}[X] := \mathrm{Cov}(X, X) = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right] \in \mathbb{R}^{n \times n}$$

  (variances are on the diagonal; covariances are on the off-diagonal)
- For non-random matrices $A \in \mathbb{R}^{* \times p}$ and $B \in \mathbb{R}^{* \times m}$,

$$\mathrm{Cov}(AX, BY) = A\mathrm{Cov}(X, Y)B^\top.$$

- For a constant vector $b \in \mathbb{R}^*$,

$$\mathrm{Var}[AX + b] = A\mathrm{Var}[X]\, A^\top$$

- $\mathrm{Var}[X]$ is positive semi-definite because

$$0 \le \mathrm{Var}\left[c^\top X\right] = c^\top \mathrm{Var}[X]\, c, \qquad c \in \mathbb{R}^n$$

Conditional Expectation

Let X and Y be RVs with finite second moments.

- Conceptually, the conditional expectation $\mathbb{E}[Y|X]$ of Y given X is the expected value of the distribution of Y conditioned on the value of X. Since X is a RV, so does $\mathbb{E}[Y|X]$
- For Y and X with a joint density $f_{X,Y}$, define

$$e_Y(x) := \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} dy$$

  The conditional expectation of Y given X is the RV $\mathbb{E}[Y|X] := e_Y(X)$
- More generally, let $\mathcal{H}$ be the smallest σ-field generated by the events $\{X \leq a\}$, $a \in \mathbb{R}$. Then $\mathbb{E}[Y|X]$ is any RV satisfying

$$\mathbb{E}[\mathbf{1}_A \mathbb{E}[Y|X]] = \mathbb{E}[\mathbf{1}_A Y], \qquad \forall A \in \mathcal{H}$$

Properties of the Conditional Expectation

Let X, Y, and Z be RVs.

- $\mathbb{E}[aY + Z|X] = a\mathbb{E}[Y|X] + \mathbb{E}[Z|X]$, $a \in \mathbb{R}$
- $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$
- $\mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|X,Z]|Z]$
- If $Y = g(X)$, then $\mathbb{E}[Y|X] = \mathbb{E}[g(X)|X] = g(X) = Y$ (Y is treated as a deterministic object under the conditional expectation)
- Law of total variance for Y with $\mathbb{E}\left[Y^2\right] < \infty$:
$$\mathrm{Var}[Y] = \mathrm{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathrm{Var}[Y|X]],$$
where $\mathrm{Var}[Y|X] := \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2 |X\right]$ (the variance under the law of Y conditioned on X)
- For Y with $\mathbb{E}\left[Y^2\right] < \infty$:
$$\mathbb{E}[Y|X] \in \arg \min_{g\,:\,g(X) \text{ is a RV}} \mathbb{E}\left[(Y - g(X))^2\right]$$
(for our purposes, this can also serves as the definition of $\mathbb{E}[Y|X]$)

Independence

- Two RVs X and Y are **independent** iff

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \qquad x, y \in \mathbb{R}$$

- Two random vectors X and Y are **independent**, iff for every measurable functions $g : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^m \to \mathbb{R}$, $g(X)$ and $h(Y)$ are indepdendent RVs
- If X and Y are independent (RVs or vectors):
  - $f_{X|Y} = f_X$ and $f_{Y|X} = f_Y$
  - $\mathbb{E}[X|Y] = \mathbb{E}[X]$ and $\mathbb{E}[Y|X] = \mathbb{E}[Y]$
  - $\mathrm{Cov}(X, Y) = 0$

The Normal Distribution

The Normal Distribution

- PDF and CDF functions of the **standard normal distribution** $Z \sim \mathcal{N}(0,1)$

$$\phi(z) := f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad \Phi(z) := F_Z(z) = \int_{-\infty}^{z} \phi(x)dx$$

- The PDF and CDF of the normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ are

$$f_X(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), \qquad F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

(σ is always assumed to be the non-negative root of σ^2)

- If $Z \sim \mathcal{N}(0,1)$, then $\sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$
- It is sometimes useful to define $\mathcal{N}(\mu, 0)$ as a point mass distribution at μ:

$$X \sim \mathcal{N}(\mu, 0) \Leftrightarrow \Pr(X \leq x) = \mathbf{1}_{x \geq \mu}.$$

Namely, $X \sim \mathcal{N}(\mu, 0)$ is the constant μ with probability one.

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a sequence of identically, independently distributed (iid) RVs with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2 < \infty$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2). \tag{1}$$

Convergence in distribution (indicated by $\xrightarrow{D}$) means pointwise convergence to a CDF, excluding points of discontinuity. In our case, (1) says that

$$\lim_{n\to\infty} \Pr\left[\sqrt{n}(\bar{X}_n - \mu) \leq z\right] = \lim_{n\to\infty} \Pr\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq \frac{z}{\sigma}\right] = \Phi\left(\frac{z}{\sigma}\right)$$

for all $z \in \mathbb{R}$.

Many other versions of the CLT exist to cover different assumptions such dependency among the RVs and/or non-identically distributed RVs.

Chi-squared distribution χ^2

- Let $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathcal{N}(0,1)$. The distribution χ_k^2 is defined as

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$$

  where k is called the number of **degrees of freedom**

- PDF:

$$f(x;\, k) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \mathbf{1}_{x \geq 0}$$

- If $X \sim \chi_k^2$, then

$$\mathbb{E}\left[X\right] = k, \quad \mathrm{Var}\left[X\right] = 2k$$

- Fun fact:

$$\mathrm{Var}\left[\sqrt{\sum_{i=1}^{k} Z_i^2}\right] = O(1)$$

t-distribution

- Suppose that $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_k$, X and Z independent. Then
$$\frac{Z}{\sqrt{\frac{X}{k}}} \sim t_k$$
where k is called the number of **degrees of freedom**
- PDF:
$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$
- If $Y \sim t_k$, then
$$\mathbb{E}[Y] = 0, \qquad \mathrm{Var}[Y] = \frac{k}{k-2}, \quad k \geq 3$$
- The t-distribution converges to the normal distribution as $k \to \infty$. It has heavier tails than that of the normal distribution.

F-distribution

- The normalized ratio of two Chisquared distribution:

$$F_{d_1,d_2} := \frac{\frac{1}{d_1}\chi^2_{d_1}}{\frac{1}{d_2}\chi^2_{d_2}}$$

- PDF:

$$f(x; d_1, d_2) = \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1} \left(1 + \frac{d_1}{d_2}x\right)^{-(d_1+d_2)/2}$$

- If $X \sim F_{n_1,n_2}$, then

$$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}, \qquad d_2 > 2$$

The Multivariate Normal

- Consider a matrix $A \in \mathbb{R}^{n \times m}$, a vector $\mu \in \mathbb{R}^m$ and the random vector

$$Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}^\top, \qquad z_i \stackrel{iid}{\sim} \mathcal{N}(0,1)$$

- The random vector $Y = AZ + \mu$ has an m-dimensional multivariate normal distribution with mean μ and variance-covariance matrix $\Sigma = AA^\top$:

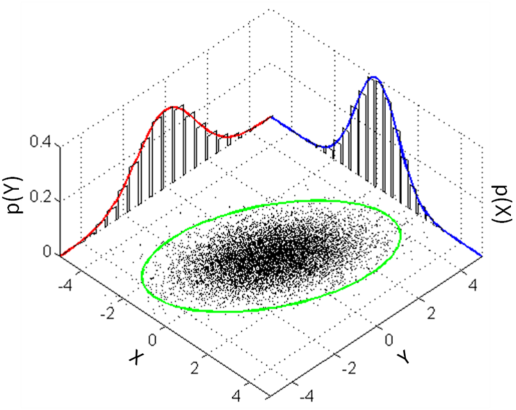
$$Y \sim \mathcal{N}(\mu, \Sigma)$$

- If Σ is invertible, then the density of Y is

$$f_Y(y) = \frac{1}{(2\pi)^{m/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right)$$

(here $y = (y_1, \ldots, y_m)$)

Depiction



(figure from Wikipedia)

Multivariate CLT

Let $X_1, X_2, \ldots, X_n$ be a sequence of iid random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_1] = \mu \in \mathbb{R}^d$ and $\mathrm{Var}[X_1] = \Sigma \in \mathbb{R}^{d \times d}$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \Sigma), \tag{2}$$

where

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \in \mathbb{R}^d$$

In our case, (2) says that

$$\begin{aligned} &\lim_{n\to\infty} \Pr\left[\sqrt{n}(\bar{X}_{n1} - \mu_1) \le z_1, \ldots, \sqrt{n}(\bar{X}_{nd} - \mu_d) \le z_d\right] \\ &\qquad = \frac{1}{(2\pi)^{d/2}|\Sigma|} \exp\left(-\frac{1}{2}\left((z-\mu)^\top \Sigma^{-1}(z-\mu)\right)^2\right) \end{aligned}$$

for all $(z_1, \ldots, z_d) \in \mathbb{R}^d$.

Linear Transformations of Normal RVs

An affine transformation of a normal vector is a normal vector

- If $Y \sim \mathcal{N}(\mu, \Sigma)$ and $U = BY + \eta$, then

$$U \sim \mathcal{N}\left(B\mu + \eta, B\Sigma B^\top\right)$$

- For example: $Z_1, Z_2 \overset{iid}{\sim} \mathcal{N}(0, 1)$, then

$$a_1 Z_1 + a_2 Z_2 \sim \mathcal{N}(0, a_1^2 + a_2^2)$$

and

$$\begin{bmatrix} a_{11} Z_1 + a_{12} Z_2 \\ a_{21} Z_1 + a_{22} Z_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} a_{11}^2 + a_{12}^2 & a_{11} a_{21} + a_{12} a_{22} \\ a_{11} a_{21} + a_{12} a_{22} & a_{21}^2 + a_{22}^2 \end{bmatrix}\right)$$

Quadratic Forms

Suppose that $A \in \mathbb{R}^{n \times n}$. The **quadratic form** associated with A is the scalar

$$x^\top A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

We have $x^\top A x = x^\top (A/2 + A^\top/2) x$, hence we can assume that A is symmetric.

Why do we care about Quadratic Forms?

- The variance estimate $\hat{\sigma}^2$ of the sample $y_1, \dots, y_n$ is a quadratic form. Indeed,

$$\hat{\sigma}^2 \propto \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

  In matrix notation:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}^\top \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & & \vdots \\ \vdots & & \ddots & \vdots \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & 1-\frac{1}{n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} y_i(y_i - \bar{y})$$

  whereas

$$\sum_{i=1}^{n} y_i(y_i - \bar{y}) = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- R^2 is of the form:

$$R^2 = 1 - \frac{y^\top A_1 y}{y^\top A_2 y}$$

Normal Quadratic Forms

Suppose $Y \sim \mathcal{N}(\mu, \Sigma)$ and that $\Sigma^{-1} \in \mathbb{R}^{n \times n}$ exists. Then

$$(Y - \mu)^\top \Sigma^{-1} (Y - \mu) \sim \chi^2_n$$

Why? eigenvalue decomposition:

- Because Σ is positive definite, we can write

$$\Sigma = P^\top \Lambda P, \qquad P^\top P = I_n, \qquad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \quad \lambda_j > 0.$$

- Define $Z = \Lambda^{-1/2} P (Y - \mu)$. Then

$$\begin{aligned} Z \sim \mathcal{N}\left(0, \Lambda^{-1/2} P \Sigma P^\top \Lambda^{-1/2}\right) &= \mathcal{N}\left(0, \Lambda^{-1/2} P P^\top \Lambda P P^\top \Lambda^{-1/2}\right) \\ &= \mathcal{N}(0, I_n) \end{aligned}$$

  It follows that $Z_1, \ldots, Z_n \overset{iid}{\sim} \mathcal{N}(0, 1)$, so

$$(Y - \mu)^\top \Sigma^{-1} (Y - \mu) = Z^\top Z = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$$

Rotation

- Suppose that:
  - $Z \sim \mathcal{N}(0, \sigma^2 I)$
  - Q is orthogonal ($Q^\top Q = I$)
- Then

$$Y = QZ \sim \mathcal{N}(0, \sigma^2 I)$$

An isotropic normal distribution is invariant under rotations

Independence

- Suppose that we partition $Y \sim \mathcal{N}(\mu, \Sigma)$ in two:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{bmatrix} \right)$$

  Y_1 and Y_2 are independent if $\mathrm{Cov}\,(Y_1, Y_2) = \Sigma_{12} = 0$ (when Σ is invertible, the proof of is from the multivariate normal density)

- For normal RVs: uncorrelatedness implies independence

Independence (cont'd)

- Suppose that $Y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.
- Define:

$$\begin{bmatrix} \bar{Y} \\ Y_1 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix} := \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ & I_n - \frac{1}{n} J_n & \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \text{where} \quad J_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \ddots & & 1 \\ \vdots & & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

- Using that $AY \sim \mathcal{N}(A\mu, \sigma^2 AA^\top)$,

$$\begin{bmatrix} \bar{Y} \\ Y_1 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & I_n - \frac{1}{n} J_n \end{bmatrix} \right)$$

  so that $\bar{Y}$ is independent of $Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}$.
- Consequently

$$\frac{(\bar{Y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{1}{n-1} \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2}}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}} \sim t_{n-1}$$

- With the normal distribution you can mine the data twice and get independent entities: one for the numerator and one for the denominator

Conditional Distributions

- Suppose that we know Y_1 in:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{bmatrix} \right)$$

What is $\mathcal{L}(Y_2|Y_1)$? (the conditional distribution of Y_2 given Y_1)

- It is Gaussian with mean

$$\mathbb{E}\left[Y_2|Y_1\right] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1)$$

and variance

$$\mathrm{Var}\left[Y_2|Y_1\right] = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}$$

Conditional Distributions (cont'd)

- In particular, if

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \right)$$

then the conditional distribution of Y given X is

$$\mathcal{L}(Y|X) = \mathcal{N}\left(\mu_y + \rho\sigma_y \frac{X - \mu_x}{\sigma_x}, \sigma_y^2(1 - \rho^2) \right)$$

- Observations:
  - The conditional mean is linear in X
  - $\Delta = (X - \mu_x)/\sigma_x$ is the number of standard deviations X is from μ_x. ρ determines the determines the relationship between Xs and Ys standard deviations.
  - The variance is independent of X

When $[Y, X]^\top$ is multivariate normal, then $\mathrm{Var}[Y|X = x]$ does not depend on which exact x was observed. Observing $X = x$ shifts the expected value of Y by a linear function of x but makes a variance change (usually a reduction) that is independent of x.