Distribution Theory

Question: What is distribution theory?
Answer: How to compute the "distribution" of an estimator, test or other statistic, T:

- Find $P(T \leq t)$, the Cumulative Distribution Function (CDF) of T.
- Find $f_T(t)$ the density of T.
- Say something like "T is $N(12, 25)$" or "T is Binomial(100,0.75)" or other named distribution. (Possible distributions include Normal, Binomial, t, Bernoulli, F, χ^2, Gamma, Geometric, Negative Binomial, Poisson, Weibull, Logistic ...)
- Find $\mathrm{E}(T)$, $\mathrm{Var}(T)$ or other "moments" of T

In this course we:

- do distribution theory when $\epsilon_i \sim N(0, \sigma^2)$
- discuss "what if the errors, ϵ_i are not normal?"
- omit proofs.

Standard normal distribution

- Standard normal density is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty$$

- We say that $Z \sim N(0,1)$ if the density of Z is standard normal:
- Reminder: if X has density $f(x)$ then

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

- Moments of Z are

$$\mathrm{E}(Z) = \int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi}} e^{-z^2/2} dz = 0$$

$$\mathrm{Var}(Z) = \mathrm{E}(Z^2) = 1.$$

General univariate normal distribution

- **Definition**: If $Z \sim N(0, 1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.
- Moments:

$$\begin{aligned} \mathrm{E}(X) &= \mu + \sigma \mathrm{E}(Z) &= \mu \\ \mathrm{Var}(X) &= \sigma^2 \mathrm{Var}(Z) &= \sigma^2 \end{aligned}$$

Multivariate normal distribution

- **Definition**: If $Z_1, \ldots, Z_n$ are independent $N(0,1)$ then

$$Z = \begin{bmatrix} Z_1 & \cdots & Z_n \end{bmatrix}^T \sim MVN_n(0, I)$$

- We say that the vector Z has a standard n-dimensional multivariate normal distribution.
- We can define $\mathrm{E}(Z)$ and $\mathrm{Var}(Z)$ for vectors like Z as follows: If X is a random vector of length n, say $X^T = [X_1 \cdots X_n]$ then

$$\mu_X \equiv \mathrm{E}(X) = \begin{bmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{bmatrix}$$

and $\mathrm{Var}(X)$ is an $n \times n$ matrix

$$\mathrm{E}\left[(X - \mu_X)(X - \mu_X)^T\right]$$

Variance Covariance Matrices

- Note that the ijth entry of $(X - \mu_X)(X - \mu_X)^T$ is

$$(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))$$

- **Definition**:

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{E}\left[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))\right] \\ &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j)\end{aligned}$$

- **Definition**: If M is a matrix then $\mathrm{E}(M)$ is a matrix whose ijth entry is $\mathrm{E}(M_{ij})$.
- So $\mathrm{Var}(X)$ has ijth entry $\mathrm{Cov}(X_i, X_j)$ and diagonal entries $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$.

Standard Multivariate Normal Moments

Now suppose $Z \sim MVN_n(o, I)$. Then

$$\mathrm{E}(Z) = \begin{bmatrix} \mathrm{E}(Z_1) \\ \vdots \\ \mathrm{E}(Z_n) \end{bmatrix} = \mathbf{0}_n$$

The $\mathbf{0}_n$ matches the 0 in $MVN_n(0, I)$.
Next we compute the variance of Z. Note that
$\mathrm{E}\left[(Z-0)(Z-0)^T\right]$ has ijth entry

$$\mathrm{E}(Z_i Z_j) = \begin{cases} 0 & i \neq j \quad \text{(independence)} \\ \mathrm{E}(Z_i^2) = 1 & i = j \end{cases}$$

So

$$\mathrm{Var}(Z) = I_{n\times n}$$

the $n \times n$ identity matrix.

General Multivariate Normal

Now suppose that

$$X = AZ + \mu$$

where

- A is an $m \times n$ matrix of constants.
- μ is a vector in R^m
- $Z \sim MVN_n(0, I)$

Then we say that X has a $MVN_m(\mu, AA^T)$ distribution.
Now $\mathrm{E}(X) = \mathrm{E}(AX + \mu)$ has ith component

$$\begin{aligned}
\mathrm{E}[(AZ)_i] + \mu_i &= \mathrm{E}(\sum_j A_{ij} Z_j) + \mu_i \\
&= \sum_j A_{ij} \mathrm{E}(Z_j) + \mu_i \\
&= \mu_i
\end{aligned}$$

Variance Covariance of MVN

Moreover,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(\left[(X-\mu)(X-\mu)^T\right] \\
&= \mathrm{E}[(AZ)(AZ)^T] \\
&= \mathrm{E}[AZZ^T A^T] \\
&= A\mathrm{E}[ZZ^T]A^T \\
&= AIA^T \\
&= AA^T
\end{aligned}$$

Thus $X \sim MVN(\mu, \Sigma)$ means that

- $\mathrm{E}(X) = \mu$
- $\mathrm{Var}(X) = \Sigma$
- X is "normal".

Things to notice along the way

1. $\mathrm{E}(AX+b) = A\mathrm{E}(X) + b$ when $A_{m\times n}$, $X_{n\times 1}$ $b_{m\times 1}$ and A and b are constant.
2. $\mathrm{E}(AMB) = A\mathrm{E}(M)B$ whenever A, B and M are matrices whose dimensions make the multiplication possible and A and B are non-random constant matrices while M is a random matrix.
3. $\mathrm{Var}(AX+b) = A\mathrm{Var}(X)A^T$ where A and X are as in 1). The notation $\mathrm{Cov}(X)$ is sometimes used for $\mathrm{Var}(X)$. This matrix is called the variance-covariance matrix of X.

Application to Least Squares

The following do not use the normal assumption:

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T Y \\
&= (X^T X)^{-1} X^T (X\beta + \epsilon) \\
&= (X^T X)^{-1} (X^T X)\beta + (X^T X)^{-1} X^T \epsilon \\
&= \beta + (X^T X)^{-1} X^T \epsilon \\
\mathrm{E}(\hat{\beta}) &= \beta + (X^T X)^{-1} X^T \mathrm{E}(\epsilon) \\
&= \beta
\end{aligned}$$

So $\hat{\beta}$ is **unbiased**.

Variance of Least Squares Estimator

$$\begin{aligned}
\mathrm{Var}(\hat\beta) &= \mathrm{E}[(\hat\beta-\beta)(\hat\beta-\beta)^T] \\
&= (X^TX)^{-1}X^T\mathrm{E}(\epsilon\epsilon^T)\left((X^TX)^{-1}X^T\right)^T \\
&= (X^TX)^{-1}X^T\sigma^2 I\left((X^TX)^{-1}X^T\right)^T \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

Distribution Theory for Least Squares

Assume:

$$\mathrm{E}(\epsilon_i) = 0 \qquad \text{and} \qquad Y = X\beta + \epsilon$$

Then

1. $\mathrm{E}(\hat\beta) = \beta$
2. $\mathrm{E}(\hat\mu) = X\mathrm{E}(\hat\beta) = X\beta = \mu$
3. $\hat\epsilon = (I - X(X^TX)^{-1}X^T)\epsilon \equiv M\epsilon$ where $M = I - X(X^TX)^{-1}X^T$.
4. $\mathrm{E}(\hat\epsilon) = M\mathrm{E}(\epsilon) = 0$

Homoscedastic errors

Define: $H = X(X^TX)^{-1}X^T$, the **hat** matrix so that $M = I - H$.
If also

$$\mathrm{Var}(\epsilon) = \sigma^2 I$$

(as will be the case for instance if the ϵ_i are iid) then

1. $\mathrm{Var}(\hat\beta) = \sigma^2(X^TX)^{-1}$
2. $\mathrm{Var}(\hat\mu) = \sigma^2 X(X^TX)^{-1}X^T = \sigma^2 H$
3. $\mathrm{Var}(\hat\epsilon) = M\mathrm{Var}(\epsilon)M^T = \sigma^2 MM^T$

Normal errors

If also $\epsilon_i \sim N(0, \sigma^2)$ are independent then

$$\begin{aligned}
\hat\beta &\sim MVN(\beta, \sigma^2 (X^T X)^{-1}) \\
\hat\mu &\sim MVN(\mu, \sigma^2 H) \\
\mathrm{Var}(\hat\epsilon) &= \sigma^2 M M^T \\
&= \sigma^2 M
\end{aligned}$$

Definition: A matrix Q is idempotent if

$$QQ \equiv Q^2 = Q$$

So What?

1. Distribution theory of Sums of Squares in ANOVA tables uses this.
2. F tests for hypotheses about parameters are justified using these matrix ideas.
3. t-tests, and confidence intervals for $c^T\beta$ (where c is a vector of length p can be derived using these ideas.

Estimation of σ

Based on the error sum of squares defined by

$$\begin{aligned} \text{ESS} &= ||\hat{\epsilon}||^2 \\ &= \sum \hat{\epsilon}_i^2 \\ &= (M\epsilon)^T(M\epsilon) \\ &= \epsilon^T M^T M\epsilon \\ &= \epsilon^T M\epsilon \end{aligned}$$

Fact

$$\mathrm{E}[\mathrm{ESS}] = \mathrm{E}(\epsilon^T M\epsilon) = \sigma^2 \sum_i M_{ii} = \sigma^2 \mathrm{trace}(M)$$

Counting Degrees of Freedom

Definition: The trace of a square matrix Q is defined by

$$\mathrm{trace}(Q) = \sum_i Q_{ii}$$

Fact:

$$\begin{aligned}
\mathrm{trace}(M) &= \mathrm{trace}(I - H) \\
&= \mathrm{trace}(I) - \mathrm{trace}(H) \\
&= n - \mathrm{trace}(\underbrace{X}_{A}\underbrace{(X^T X)^{-1} X^T}_{B}) \\
&= n - \mathrm{trace}(\underbrace{(X^T X)^{-1} X^T X}_{I_{p\times p}}) \\
&= n - \mathrm{trace}(I_{p\times p}) \\
&= n - p
\end{aligned}$$

Notice that p is the number of columns of X **including** the column of 1's if present.

Summary of result

$$\mathrm{E}\left[\frac{\mathrm{ESS}}{n-p}\right] = \sigma^2$$

So the Mean Squared Error, $ESS/(n-p)$ is an unbiased estimate of σ^2.

Inference for linear combinations of entries in β

Examples of linear combinations:

- Confidence intervals and tests for μ_x, the mean of Y corresponding to a particular value x of the covariate. (See "Polynomial Regression" page for example.)

$$\hat{\mu}_x = x_1\hat{\beta}_1 + \cdots + x_p\hat{\beta}_p = x^t\hat{\beta}$$

- Confidence intervals and tests for $\beta_k = [0, \ldots, 0, 1, 0, \ldots, 0]\beta$ where the 1 is in position k.

Basic Ingredients

1. $\mathrm{E}(x^t\hat{\beta}) = x^T\beta = \mu_x$.
2. $\mathrm{Var}(x^t\hat{\beta}) = x^T\mathrm{Var}(\hat{\beta})x = \sigma^2 x^T(X^TX)^{-1}x$
3. $\hat{\sigma}^2 = \mathrm{ESS}/(n-p)$ is consistent (that is, converges to σ^2 as $n \to \infty$).
4. If $\epsilon_i \sim N(0,\sigma^2)$ **or** n is large then

$$\hat{\beta} \sim MVN(\beta, \sigma^2(X^TX)^{-1})$$

5. If as in 4 then

$$\hat{\mu}_x \sim MVN(\mu_x, \sigma^2 x^T(X^TX)^{-1}x)$$

6. If $\epsilon_i \sim N(0,\sigma^2)$ then $\mathrm{ESS}/\sigma^2 \sim \chi^2_{n-p}$
7. If $\epsilon_i \sim N(0,\sigma^2)$ then $\hat{\beta}$ is independent of ESS.

More points

8. If $\epsilon_i \sim N(0, \sigma^2)$ then

$$\frac{\hat{\mu}_x - \mu_x}{\hat{\sigma}\sqrt{x^T(X^TX)^{-1}x}} \sim t_{n-p}$$

9. A confidence interval for $x^T\beta$ is

$$x^t\hat{\beta} \pm t_{\alpha/2,n-p}\sqrt{\frac{\text{ESS}}{n-p}}\sqrt{x^T(X^TX)^{-1}x}$$

10. The confidence interval in 9 is justified either by
    - Normal errors OR
    - large n