FALL 2013
Stat 8003: Statistical Methods I
Lecture 10

Jichun Xie

1 Sample Size and Power Analysis for Normal Distribution

Example: We are interested in determining whether the mean volume of fluid delivered to patients during a particular type of neuro-surgery is at least 50ml greater than 1500ml. If so then the surgeons need to think about modifying the procedures to reduce the large fluid volumes. Our null hypothesis is that the mean volume, μ, is 1500 ml. We cannot rule out $\mu < 1500$. We believe σ^2 is roughly 10,000 ml^2.

Questions of Interest:

1. What power will we have to detect a difference in the mean of 50 ml if we use a sample size of 10 patients?

2. How many patients will we need to detect a difference of 50 ml with at least 80% power?

3. Setup:

$$\begin{aligned} \mathrm{H}_0 &: \mu = \mu_0 \\ \mathrm{H}_1 &: \mu \neq \mu_0, \end{aligned}$$

where μ is the mean volume of fluid. The type I error rate is $\alpha = 0.05$.

4. Let $Y_1, \ldots, Y_n$ represent the sample of volumes. The test-statistic will be the Z-statistic (Sometimes it is also called T-statistic, but note it does not mean it follows t-distribution since $Y_1, \ldots, Y_n$ can be non-normal). For the purpose of sample size/power calculations we use:

$$Z = \frac{\bar{Y} - \mu_0}{\sqrt{\hat{\sigma}^2/n}}.$$

Under H_0, asymptotically $Z \sim \mathrm{N}(0,1)$.

5. Under the null

$$\mathbb{P}\left(|Z| > \left|\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right| \mid \mathrm{H}_0 \text{ is true }\right) = \alpha.$$

6. Under the alternative we want

$$\mathbb{P}\left(|Z| > \left|\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right| \mid \mathrm{H}_1 \text{ is true }\right) = 1-\beta.$$

$$\mathbb{P}\left(\frac{\bar{Y}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} > \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) + \mathbb{P}\left(\frac{\bar{Y}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} < \Phi^{-1}\left(\frac{\alpha}{2}\right)\right) = 1-\beta.$$

7. Let μ_1 be the mean under the alternative. WLOG, suppose $\mu_1 > \mu_0$.

$$\mathbb{P}\left(\frac{\bar{Y}-\mu_1}{\sqrt{\frac{\sigma^2}{n}}} > \Phi^{-1}\left(1-\frac{\alpha}{2}\right) + \frac{\mu_0-\mu_1}{\sqrt{\frac{\sigma^2}{n}}}\right) + \mathbb{P}\left(\frac{\bar{Y}-\mu_1}{\sqrt{\frac{\sigma^2}{n}}} < \Phi^{-1}\left(\frac{\alpha}{2}\right) + \frac{\mu_0-\mu_1}{\sqrt{\frac{\sigma^2}{n}}}\right) = 1-\beta.$$

8. But since $\mu_1 > \mu_0$, the second term is small, and we usually ignore it. Why?

9. What is the distribution fo the first term in (7) when H_1 is true?

$$1-\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) + \frac{\mu_0-\mu_1}{\sqrt{\frac{\sigma^2}{n}}}\right) \approx 1-\beta.$$

10. The equation in (9) gives us the power. What about the sample size?

$$\begin{aligned}
\Phi^{-1}\left(1-\frac{\alpha}{2}\right) + \frac{\mu_0-\mu_1}{\sqrt{\frac{\sigma^2}{n}}} &= \Phi^{-1}(\beta) \\
\frac{\mu_0-\mu_1}{\sqrt{\frac{\sigma^2}{n}}} &= \Phi^{-1}(\beta) - \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \\
\frac{(\mu_0-\mu_1)^2}{[\Phi^{-1}(\beta)-\Phi^{-1}(1-\frac{\alpha}{2})]^2} &= \frac{\sigma^2}{n} \\
n = \frac{\sigma^2[\Phi^{-1}(\beta)-\Phi^{-1}(1-\frac{\alpha}{2})]^2}{(\mu_0-\mu_1)^2}
\end{aligned}$$

11. Back to the example: $\sigma^2 = 100^2(\mu_0-\mu_1) = -50$, $\Phi^{-1}(\beta) = -0.84$, $\Phi^{-1}(1-\frac{\alpha}{2}) = 1.96$,

$$n = 31.36.$$

Conclude that we need at least 32 patients to have at least 80% power to detect an increase of 50 ml.

12. Suppose that we were wrong and the difference is actually a decrease in the volume by 50 ml. What would the power be?

13. What if we use 10 patients? What is the power?

$$1 - \beta = 1 - \Phi\left(1.96 + \frac{-50}{\sqrt{\frac{100^2}{10}}}\right) = 0.36.$$

14. Suppose we carry out the experiment with 10 patients and don't reject the null. What would we conclude from the experiment?

2 Linear Regression

2.1 Introduction

In lots of the real problems, the data can be formulated to a response variable Y and some covariates X.

Y	X
dependent variable	independent variable
response variable	explanatory variable
outcome variable	covariate
	predictor

Linear regression can handle the case where Y is continuous. In next semester, we are going to discuss logistic regression and Poisson regression. These two models can handle the case where Y is categorical or count.

Purpose of "Regression": quantify the magnitude of the association/relationship between Y and X.

Example. In a phase-II clinical trial, the researchers are interested in the effect of a new cholesterol lowering drug. The patients are randomized to different drug-dose group and the cholesterol levels before and after taking the drug are measured for each patient. The researchers want to know:

1. Is the new drug has any effect of lowering cholesterols? If there is some effect, is the effect strong or weak? (Estimation)
2. Given a new dose X, what would Y be? (Prediction)

Suppose

$$\begin{aligned} X_i &= \text{dose of the drug the } i\text{-th patent took} \\ Y_i &= \text{the difference of the cholesterol level of the } i\text{-th patient} \end{aligned}$$

We can view (X_i, Y_i) are n *i.i.d.* realization of the random variable X and Y.

How to study the relationship between Y and X?

Useful plots:

1. Scatterplot: for continuous Y and X (See Figure 1)

   $$\text{Plot } Y_i \textit{ vs. } X_i \text{ for all } i.$$

   Features:

- Overal shape: linear/nonlinear
- Spread
- Isolated points

2. Boxplot: for continuous Y but categorical X.

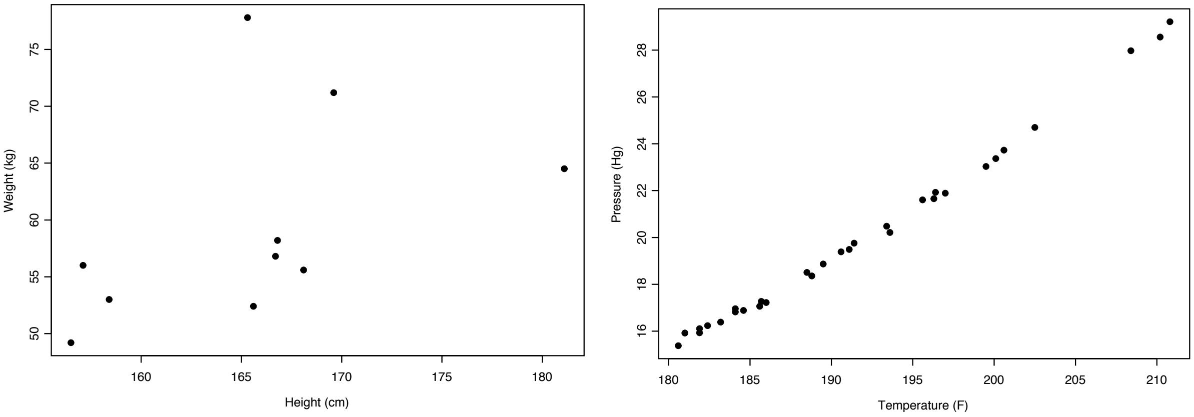

Figure 1: Scatterplot examples

Sample Conditional Mean and Population Conditional Mean:

- Avg$(Y \mid X)$, "$|$" means given.
- $\mathtt{E}(Y \mid X)$, conditional expectation.

2.2 Relationship in Statistical Terms

Correlation $\rho(X, Y)$ or ρ_{XY}: measures the linear association between two r.v. Y and X.

$$\rho_{XY} = \frac{\mathtt{Cov}(X, Y)}{\sqrt{\mathtt{Var}(X) \cdot \mathtt{Var}(Y)}},$$

where

$$\mathtt{Cov}(X, Y) = \mathbb{E}\{(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\}, \quad \mathtt{Var}(Y) = \mathbb{E}\{(Y - \mathbb{E}(Y))^2\} = \sigma_Y^2$$

Note that $-1 \leq \rho(X, Y) \leq 1$. Why?

Consider $Z_1 = \frac{Y}{\sigma_Y} + \frac{X}{\sigma_X}$ and $Z_2 = \frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}$.

$$\begin{aligned}
\texttt{Var}\,(Z_1) &= \texttt{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\
&= \texttt{Var}\left(\frac{X}{\sigma_X}\right) + \texttt{Var}\left(\frac{Y}{\sigma_Y}\right) + 2\texttt{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\
&= 1 + 1 + 2\rho(X,Y).
\end{aligned}$$

Since $\texttt{Var}(Z_1) \geq 0$, $1 + 1 + 2\rho(X,Y) \geq 0$, and then $\rho(X,Y) \geq -1$.

Similarly by calculating $\texttt{Var}(Z_2)$ and use the fact $\texttt{Var}(Z_2) \geq 0$, we can show $\rho(X,Y) \leq 1$.

Here are some other facts about $\rho_{X,Y}$:

- When $\rho_{XY} = 0$,
$$\texttt{Var}\left(\frac{Y}{\sigma_Y} + \frac{X}{\sigma_X}\right) = \texttt{Var}\left(\frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}\right).$$

- When $\rho_{XY} > 0$,
$$\texttt{Var}\left(\frac{Y}{\sigma_Y} + \frac{X}{\sigma_X}\right) > \texttt{Var}\left(\frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}\right).$$

- When $\rho_{XY} = 1$,
$$\texttt{Var}\left(\frac{Y}{\sigma_Y} - \frac{X}{\sigma_X}\right) = 0.$$

Some extra notes about $\rho_{X,Y}$:

- If X and Y are independent ($X \perp\!\!\!\perp Y$), then $\rho_{X,Y} = 0$.

- However $\rho_{X,Y} = 0$ doesn't necessarily lead to $X \perp\!\!\!\perp Y$.

- However, if X and Y follow normal distribution, then $\rho_{X,Y} = 0$ leads to $X \perp\!\!\!\perp Y$.

- ρ_{XY} is a measure of linear association, it is **NOT** a measure of causality.

Sample correlation: After obtaining the data (X_i, Y_i), $i = 1, \ldots, n$. We can define the sample correlation (Pearson's correlation) as:

$$\begin{aligned} r_{XY} &= \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 \cdot \frac{1}{n}\sum i = 1^n (X_i - \bar{X})^2}} \\ &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \cdot \sum_{i=1}^n (X_i - \bar{X})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \end{aligned}$$

where $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$, $S_{yy} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $S_{xy} = \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})$.

2.3 Basic Regression Model

$$\begin{aligned} Y &= \text{Systematic component} + \text{Random component} \\ &= \text{Fit} + \text{Error or ''Residuals''} \\ &= f(X;\beta) + \epsilon \end{aligned}$$

1. Fit $f(X;\beta)$: estimate the parameter β. The logic is $f(X;\beta)$ should be close to Y so that the error term ϵ is small.

2. Error ϵ: described by some probability distributions.

Consider fitting a function $f(X;\beta)$ such that $\mathbb{E}(Y \mid X) = f(X;\beta)$. If $f(X;\beta) = X\beta$, then the model is called a linear model.

Simple linear regression. Suppose the data are (X_i, Y_i), $i = 1, \ldots, n$. Consider the model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ &= \text{fit}(X;\beta) + \text{Error} \end{aligned}$$

Here the systematic component $f(X;\beta) = \beta_0 + \beta_1 X_i$ is linear. The random component is ϵ_i.

Figure 2 shows an example of simple linear regression.

Remarks:

1. The model is called "simple" because there is only one covariate X.

2. There is not too much difference between the fix design and the random design. The key quantity is $\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$.

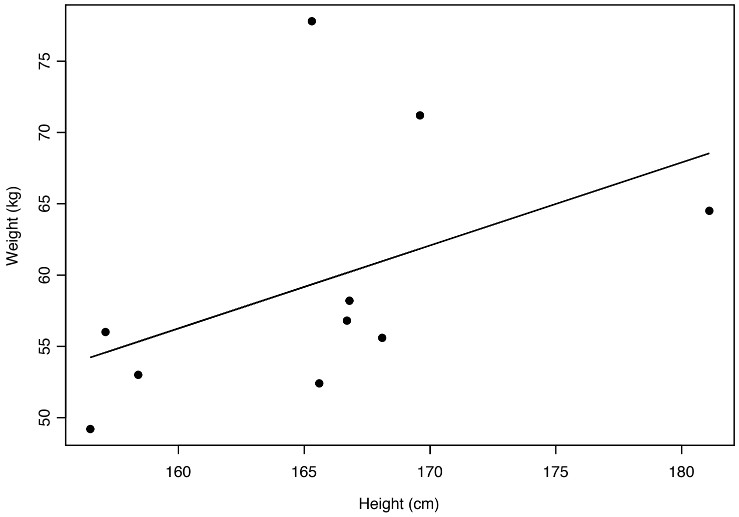

Figure 2: Simple linear regression: weight *vs.* height

Here are some assumptions about SLR:

1. Zero mean of the error: $\mathbb{E}\left(\epsilon_i\right) = 0$.
2. Constant variance of the error: $\texttt{Var}(\epsilon_i) = \sigma^2$.
3. ϵ_i's are mutually independent: $\epsilon_i \perp\!\!\!\perp \epsilon_j$, $i \neq j$.
4. ϵ_i's are normal r.v. (needed for testing).

Assumption 1 and 2 are equivalent to the following

1*. $\mathbb{E}\left(Y_i \mid X_i\right) = \beta_0 + \beta_1 X_i$.

2*. $\texttt{Var}(Y_i \mid X_i) = \sigma^2$.

If assumptions 1, 2, 3 and 4 holds, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

How to interpretate the model?

Example. Suppose X = Years of experience, Y = annual salary of an employee (in \$1000). Consider two employees with x, $x+1$ years of experience.

The expected salary for employee 1 is

$$\mathbb{E}\left(Y \mid X = x\right) = \beta_0 + \beta_1 x. \tag{1}$$

And the expected salary for employee 2 is

$$\mathbb{E}\left(Y \mid X = x+1\right) = \beta_0 + \beta_1 (x+1). \tag{2}$$

Then eq(2) - eq(1) = β_1. Therefore, β_1 can be interpreted as

- *difference* in *average* salary (Y) between employees with $x+1$ years of experience and x years of experience

- *difference (change)* in *average* Y with 1 unit *increase (change)* in X.

Interpretation about β_0: $\beta_0 = \mathbb{E}(Y \mid X = 0)$ is the expected salary for an employee with no working experience.

In other examples, β_0 can be less meaningful, *e.g.*. relationship between babies' weight (Y) and height (X).

Next we discuss the estimation of the unknown parameter β_0, β_1 and σ^2. Previously, we discussed two general types of estimators, including method of moments (MOM) and maximum likelihood estimators (MLE). For simple linear regression, we can use the above methods too. But due to the special properties of the model, we first start with another method with more intuitive explanations.

2.4 Least Square Estimates (LSE) for Simple Linear Regression

Logic: compared to the systematic terms, the error terms are very small. In other words, Y_i is very "close" to $X_i\beta_1 + \beta_0$. To measure the closeness, we can use squared error:

$$\epsilon_i^2 = \{Y_i - (X_i\beta_1 + \beta_0)\}^2.$$

Loss function:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)^2 \tag{3}$$

The estimator of β_0 and β_1 is the minimizer of eq(3):

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} L(\beta_0, \beta_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)^2 \tag{4}$$

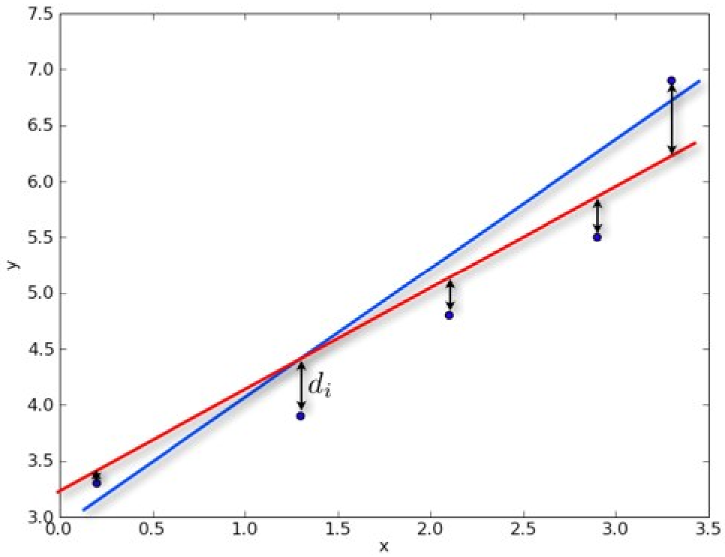

Figure 3: Residual Sum Squares

$$
\begin{aligned}
\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0\\
&\Rightarrow \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0\\
&\Rightarrow \hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x}\\
\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i=0\\
&\Rightarrow \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0\\
&\Rightarrow \sum_{i=1}^{n} x_i y_i - (\bar{y}-\hat{\beta}_1\bar{x})n\bar{x} - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0\\
&\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \hat{\beta}_1(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = 0\\
&\Rightarrow S_{xy}-\hat{\beta}_1 S_{xx}=0
\end{aligned}
$$

The last step is true since

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\
S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}
$$

Next, we discuss how to estimate σ^2.

$$\epsilon_i \sim N(0, \sigma^2) \ \textit{i.i.d.} \quad \text{and} \quad \texttt{Var}(\epsilon_i) = \sigma^2.$$

If β_0, β_1 are known, then

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i,$$

the variance can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

When β_0 and β_1 are unknown, we can replace them with $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

However, it is a biased estimator. We need to reduce the *degrees of freedom (d.f.)* since we estimate β_0 and β_1.

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{RSS}{n-2}$$

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, fitted value of y_i;

- RSS = *Residual Sum of Square* = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.