1 W1

1.1 Review.

Mean/Expectation

The **mean** or **expectation** of a continuous random variable Y is given by

$$\mathbb{E}[Y] = \int y f(y)\, dy$$

For random variables $Y_1, \ldots, Y_m$ and constants a_i, b_i for $i = 1, \ldots, m$,

$$\mathbb{E}\left[\sum_{i=1}^{m}(a_i Y_i + b_i)\right] = \sum_{i=1}^{m} a_i \mathbb{E}[Y_i] + \sum_{i=1}^{m} b_i.$$

This is called the **linearity of expectation**. For observations $y_1, \ldots, y_n$, the **sample mean** is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Variance

The **variance** of a continuous random variable Y is given by

$$\operatorname{Var}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - \mathbb{E}^2[Y]$$

- For constants $a, b \in \mathbb{R}$, $\operatorname{Var}[aY + b] = a^2 \operatorname{Var}[Y]$.
- If X and Y are *independent*, then $\operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$.

For observations $y_1, \ldots, y_n$, the **sample variance** is

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

Covariance

The **covariance** of two continuous random variables X, Y is given by

$$\operatorname{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$.
- $\operatorname{Cov}(aY + c, bX + d) = ab \cdot \operatorname{Cov}(X, Y)$.
- $\operatorname{Cov}(U + V, X + Y) = \operatorname{Cov}(U, X) + \operatorname{Cov}(U, Y) + \operatorname{Cov}(V, X) + \operatorname{Cov}(V, Y)$.
- $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)$.

For observations $(y_1, x_1), \ldots, (y_n, x_n)$, the **sample covariance** is

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}).$$

Normal Distribution

$$
\begin{aligned}
Z &\sim N(\mu, \sigma^2) \\
\mathbb{E}[Z] &= \mu \\
\mathrm{Var}(Z) &= \sigma^2
\end{aligned}
$$

For independent $Z_i \sim N(\mu_i, \sigma^2)$, $U = \sum_{i=1}^{n}(a_i Z_i + b_i)$ is normally distributed, i.e.,

$$U \sim N\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

Chi-Square Distribution

$$
\begin{aligned}
X &\sim \chi^2_\nu \quad (\nu \text{ denotes the degrees of freedom}) \\
\mathbb{E}[X] &= \nu \\
\mathrm{Var}(X) &= 2\nu.
\end{aligned}
$$

For standard normal random variables $Z_i \overset{\text{iid}}{\sim} N(0,1)$,

$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

t-Distribution

$$
\begin{aligned}
Y &\sim t_\nu \quad (\nu \text{ denotes the degrees of freedom}) \\
\mathbb{E}[Y] &= 0 \quad \text{if } \nu > 1, \text{ otherwise NA} \\
\mathrm{Var}[Y] &= 2\nu \quad \text{if } \nu > 2, \text{ otherwise } \infty
\end{aligned}
$$

For *independent* $Z \sim N(0,1)$ and $X \sim \chi^2_\nu$,

$$\frac{Z}{\sqrt{X/\nu}} \sim t_\nu.$$

1.2 Motivation: Toward Linear Regression.

- How do we characterize the relationship between x and y?
- How do we predict y given x?
- How does the mean of y change when x increases by a?

Simple Linear Regression

We can answer questions like these with **simple linear regression**.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Intuitively, we are assuming that there exists some underlying linear relationship between the covariates x and the observations y, where β_0 and β_1 are unknown:

$$y \approx \beta_0 + \beta_1 x.$$

The error term ε captures the difference between the actual y and the predicted $\beta_0 + \beta_1 x$.

Multiple Linear Regression

What if we have multiple covariates? Suppose each sample x_i has three covariates x_{i1}, x_{i2}, x_{i3}. We can generalize the simple linear regression to **multiple linear regression**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

Note that each covariate x_{ij} has a corresponding β_j parameter.

Course Outlook

This course will focus on developing multiple linear regression:

- Theoretically/mathematically: derive estimators.
- Practically: how to fit these models in R.
- How to choose and compare a model, i.e., which x_{ij} to include;
- How to evaluate the appropriateness of the model and assumptions