Review of Statistics

1 Random Variables and Key Statistics

Random Variable: A random variable is a variable that takes on different numerical values from a sample space determined by chance (probability distribution, f(x)). For example, the outcome of rolling a fair dice is a random variable having possible values of $1, \ldots, 6$ each with a chance of $\frac{1}{6}$. A random variable is *discrete* if it can assume at most a countable number of values.
Key statistics for a random variable, X:

- Expected value $\mu = E(X) = \sum_{\text{all } x} xf(x)$, for example, $\mu = \sum_{x=1}^{6} \frac{1}{6}x$ from rolling a fair dice.
- Variance: measures the level of dispersion of a random variable- average square distance to the mean.
$$\sigma^2 = V(X) = E[(X-\mu)^2] = \sum_{\text{all } x} (x-\mu)^2 f(x)$$
or
$$\sigma^2 = V(X) = E(X^2) - [E(X)]^2$$

Test hypotheses: Two-Tailed, Large-Sample Test for the Population Mean

- $H_0 : \mu = \mu_0$
  $H_1 : \mu \neq \mu_0$
- The significant level of the test: α (usually, we set $\alpha = 0.01$, 0.05, or 0.1)
- Test statistic: $z = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
- Critical points: $\pm Z_{\alpha/2}$
- The decision rule: Reject the null hypothesis if either $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$.

Example 1.1 *An insurance company executive believes that, over the last few years, the average liability insurance per board seat in companies defined as "small companies" has been \$2,000. A recent survey of small business by Growth Resources, Inc., reports that the average liability tab per board seat in their sample is \$2,700. Assume that the sample used by Growth Resources contained 100 randomly chosen small firms (as defined by their total annual gross billing) and that the sample standard deviation was \$947. Do these sampling results provide evidence to reject the executive's claim that the average liability per board seat is \$2,000, using a $\alpha = 0.01$ level of significance?*

Answer: We set $H_0 : \mu = 2000$ and $H_1 : \mu \neq 2000$. Since $\alpha = 0.01$, we have the Z statistics for the two critical points, $\pm Z_{\alpha/2} = \pm 2.575$, while the test statistic

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2700 - 2000}{947/10} = \frac{700}{94.7} = 7.39 > Z_{\alpha/2}$$

Thus, we reject the null hypothesis.

2 Measures of Association Between Two Variables

In data analysis, we sometimes want to learn the relationship between two variables, for example, does the higher temperature in July lead to higher electricity consumption? The statistics, covariance and correlation serve for that purpose. They are the building blocks of many advanced multi-variate analysis.

- sample covariance: $s_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$
- population covariance: $\sigma_{xy} = cov(x,y) = \frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{N}$

  Effect of variable scaling:
- Pearson sample correlation coefficient: $r_{xy} = \frac{s_{xy}}{s_x s_y}$ where s_x and s_y are sample standard deviations of random variables x and y respectively, and $s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$, $s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$. Note that $-1 \leq r_{xy} \leq 1$.
- Pearson population correlation coefficient: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ where σ_x and σ_y are population standard deviations of random variables x and y respectively, and $\sigma_x = \sqrt{\frac{\sum(x_i-\mu_x)^2}{N}}$, $\sigma_y = \sqrt{\frac{\sum(y_i-\mu_y)^2}{N}}$
- Graphic Interpretations:

Example 2.1 *The following data set contains 2 variables and 10 observations. For example, the data might come from a survey to 10 female respondents. Variable x represents the number of children the respondent has and variable y records the age of the respondent. We are interested in knowing whether the older generation tends to raise more children than the younger generation. Note that all respondents are either in their late stage of reproductive period or has passed that period. For survey data, we usually arrange the data in rows and columns, each row corresponding to the answers to all survey questions from a respondent and each column listing answers to one question from all respondents.*

obs.	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	*2*	*50*	*-1*	*-1*	*1*
2	*5*	*57*	*2*	*6*	*12*
3	*1*	*41*	*-2*	*-10*	*20*
4	*3*	*54*	*0*	*3*	*0*
5	*4*	*54*	*1*	*3*	*3*
6	*1*	*38*	*-2*	*-13*	*26*
7	*5*	*63*	*2*	*12*	*24*
8	*3*	*48*	*0*	*-3*	*0*
9	*4*	*59*	*1*	*8*	*8*
10	*2*	*46*	*-1*	*-5*	*5*
Sum	*30*	*510*	*0*	*0*	*99*
Average	*3*	*51*	*0*	*0*	*9.9*

Answer:

- $s_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{99}{10-1} = 11$
- $s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.4907$
- $s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.9303$
- $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.4907)(7.9303)} = 0.93$

When two variables X and Y are positively correlated, higher value of X usually comes with higher value of Y and smaller value of X is more likely to be associated with small value of Y.

3 Linear Combinations of Random Variables

We now consider multi-variate cases where there are two or more variables. Let's consider a scenario where we are creating a portfolio consisting of n individual stocks with initial capital one million dollars. It is our decision to determine the percentage of the initial capital to be invested in each stock so that certain goals can be achieved, for example, at least 10% expected daily return and no more than 15% risk (measured by standard deviation). To facilitate the decision process, we need to evaluate the portfolio expected return and risk under various alternatives. Assuming that in one alternative, we invest a_i portion of total capital in stock i, where $0 \leq a_i \leq 1$ and $\sum_{i=1}^{n} a_i = 1$, we can find the expected return and risk for this alternative if we know the expected daily return and risk for each individual stock. The expected return and risk of each individual stock can be obtained from historical data. For example, the expected daily return of stock i is the mean daily return of stock i in the past three years (or any duration in which we have data) and the expected risk of stock i is the standard deviation of the daily return in the same period. In addition to return and risk, we also need to know the covariance between any pair of stocks in the portfolio. This can again be obtained from the historical data. Once the information about the individual stock is available, the expected return and risk of the portfolio given the portfolio composition a_i, $i = 1, \ldots, n$ can be easily calculated following the following theorems. In this example, the daily return of each stock X_i, $i = 1, \ldots, n$ is a random variable and the daily return of the portfolio is also a random variable which is a linear combination of the n individual random variables ($X_p = a_1X_1 + a_2X_2 + \cdots + a_nX_n$).

Theorem 1 *Let $X_1, X_2, \ldots, X_n$ be random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ respectively. Then*

$$\begin{aligned} E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] &= a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n] \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \end{aligned} \quad (1)$$

$$Var[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 + 2\Sigma_{i=1,i<j}^{n}\Sigma_{j=1}^{n}a_ia_jCov(X_i, X_j) \quad (2)$$

where

$$\mu = E(X) = \sum_{\text{all } x} x f(x)$$

and

$$\sigma^2 = V(X) = E[(X-\mu)^2] = \sum_{\text{all } x} (x-\mu)^2 f(x)$$

Theorem 1 shows that the expected value of the linear combination of some random variables is the linear combination of the means of those variables.

Theorem 2 *Let $X_1, X_2, \ldots, X_n$ be* **independent** *random variables with means $\mu_1, \mu_2, \ldots,$ μ_n and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ respectively. Then*

$$Var[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

When two variables X_i and X_j are independent, $Cov(X_i, X_j) = 0$. Thus, the last term in the variance formula is gone.

Theorem 3 *Let $X_1, X_2, \ldots, X_n$ be* **independent identically distributed** *random variables with mean μ and variance σ^2.*

$$Var[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = [a_1^2 + a_2^2 + \cdots + a_n^2]\sigma^2$$

Example 3.1 *Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables with mean μ and variance σ^2.*

$$\begin{aligned} E[\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n] &= \tfrac{1}{n}E[X_1] + \tfrac{1}{n}E[X_2] + \cdots + \tfrac{1}{n}E[X_n] \\ &= \tfrac{1}{n}\mu + \tfrac{1}{n}\mu + \cdots + \tfrac{1}{n}\mu = n(\tfrac{1}{n}\mu) = \mu \end{aligned}$$

$$Var[\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n] = (\frac{1}{n})^2\sigma^2 + (\frac{1}{n})^2\sigma^2 + \ldots + (\frac{1}{n})^2\sigma^2 = n(\frac{1}{n})^2\sigma^2 = \frac{\sigma^2}{n}$$

Note that $\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n = \frac{X_1+X_2+\cdots+X_n}{n} = \bar{X}$. So $E[\bar{X}] = \mu$, $Var(\bar{X}) = \frac{\sigma^2}{n}$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, that is, the mean and standard deviation of the sampling distribution are μ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, respectively. It is clear that as sample size n increases, the standard deviation of the sampling distribution becomes smaller.

We now rewrite Theorem 1 in matrix forms. Let $\mathbf{m} = [\mu_1, \mu_2, \ldots, \mu_n]^T$ and $\mathbf{C}$ be the variance covariance matrix, i.e.

$$\mathbf{C} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{n,n-1} & \sigma_{nn} \end{bmatrix}$$

Outcome	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Table 1: Possible outcomes of Y

If $\mathbf{a} = [a_1, a_2, \ldots, a_n]^T$ contains the coefficients of the linear combination in Theorem 1, Equations (1) and (2) can be rewritten as

$$E(Y) = \mathbf{a}^T\mathbf{m}$$

and

$$Var(Y) = \mathbf{a}^T\mathbf{Ca}$$

where $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$.

Example 3.2 *Let X be a discrete uniformly distributed random variable with possible values $1, 2, \ldots, 6$. Find the mean and standard deviation of the mean of nine randomly chosen observations.*

$$\mu = E(X) = \Sigma_{\text{all } x} xf(x) = (1)(\frac{1}{6}) + (2)(\frac{1}{6}) + \cdots + (6)(\frac{1}{6}) = \frac{21}{6} = 3.5$$

$$Var(X) = E(X-\mu)^2 = \Sigma_{\text{all } x}(x-\mu)^2 f(x) = E(X^2) - [E(X)]^2$$

Since $E(X^2) = (1^2)(\frac{1}{6}) + (2^2)(\frac{1}{6}) + \ldots + (6^2)(\frac{1}{6}) = \frac{91}{6}$, $Var(X) = E(X^2) - [E(X)]^2 = \frac{91}{6} - (\frac{21}{6})^2 = \frac{546}{36} - \frac{441}{36} = \frac{105}{36} = 2.9167$

$$\sigma_X = 1.7$$

$$E(\bar{X}) = E(X) = \mu = 3.5$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.7}{\sqrt{9}} = .57$$

If we rolled a set of nine fair dices and averaged the number of dots on the top faces, we would expect that this average would be between 3.5- 1.96(.57) and 3.5+1.96(.57) or between 2.38 and 4.62 about 95% of the time if we believe the Central Limit Theorem applies to a sample this small.

Example 3.3 *Let X_1, X_2 be discrete uniformly distributed random variables with possible values $1, 2, \ldots, 6$. Find the mean and standard deviation of the random variable $Y = X_1 + X_2$.*

From Table 1, we have

$$\begin{aligned}\mu = \quad E(Y) = \quad & \Sigma_{\text{all } x} x f(x) = (2)(\frac{1}{36}) + (3)(\frac{2}{36}) + (4)(\frac{3}{36}) + (5)(\frac{4}{36}) + (6)(\frac{5}{36}) + (7)(\frac{6}{36}) \\ & +(8)(\frac{5}{36}) + (9)(\frac{4}{36}) + (10)(\frac{3}{36}) + (11)(\frac{2}{36}) + (12)(\frac{1}{36}) = \frac{252}{36} = 7\end{aligned}$$

This is the same as $2 \times E(X) = 2 \times 3.5$.

$$\begin{aligned}Var(Y) = \quad E(Y-\mu)^2 = \quad & \Sigma_{\text{all } y}(y-\mu)^2 f(y) = E(Y^2) - [E(Y)]^2 \\ & = [2^2(\frac{1}{36}) + 3^2(\frac{2}{36}) + 4^2(\frac{3}{36}) \\ & +5^2(\frac{4}{36}) + 6^2(\frac{5}{36}) + 7^2(\frac{6}{36}) \\ & +8^2(\frac{5}{36}) + 9^2(\frac{4}{36}) + 10^2(\frac{3}{36}) \\ & +11^2(\frac{2}{36}) + 12^2(\frac{1}{36})] - 7^2 = \frac{1974}{36} - 49 = 5.8333\end{aligned}$$

$$\sigma_Y = \sqrt{5.8333} = 2.415$$

We can also obtain it from

$$Var(Y) = Var(X) + Var(X) = 2Var(X) = 2 * 2.9167$$

Example 3.4 *If a car dealer estimated that she has a 30% of chance selling 3 cars a day, a 40% of chance selling 2 cars a day, a 20% chance of selling 1 car a day and 10% of chance with no sales in a day.*

1. *What is the expected number of cars sold by the dealer and what is the standard deviation?*
2. *If the dealer now owns 3 stores and the distribution of number of cars sold in a day is identical in all stores, what is the expected* **total** *number of cars sold in a day by the 3 stores and what is the standard deviation of the total? (assume the distribution for each store is the same as the one described for the one store case)*
3. *In the 3-store case, what is the expected value of the* **average** *number of cars sold a day from the three stores and what is the standard deviation of the average?*