Pharmacy 2012: Biostatistics

Recommended book: DeVeaux, Velleman, Bock, Vukov, Wong: Stats: Data and Models, 3rd Canadian Edition, Pearson, 2019

Population, sample, random variable, and distribution

- Based on a random sample, we wish to make inferences about the population.
- We represent the values in the population by a random variable, say X, and a probability distribution.
  - A random variable is just a symbol for the next value we will get.
  - The probability distribution can be represented by a bar graph or smooth density curve.
  - The probability usually depends on some constants, called parameters.
- Discrete random variables have probability in chunks at separated values.
  - For example the binomial random variable, X, is the number of successes in n binary trials, and has probability mass function

    $$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

    for $x = 0, \ldots, n$
- Continuous random variables can, in theory, take on values in intervals, and the probability is given by the area under a probability density curve.
  - For example, the normal random variable has a probability density with a bell shaped density curve.
  - Areas under the curve are often obtained using tables.

Problems: If Z has a standard normal density,

1. Find $P(Z \leq 1.96)$, the probability that Z is less than or equal to 1.96. Answer: .975
2. Find $P(Z > 1.96)$. Answer: 1-.975=.025
3. Find $P(Z \leq -1.96)$. Answer: .025 (same as part b, as normal distribution is symmetric about 0)
4. Find $P(-1.96 \leq Z \leq 1.96)$. Answer: .95 ($P(-1.96 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z \leq -1.96) = .975 - .025 = .95$.
5. What is the value of c for which $P(Z \leq c) = .9370$? Answer: $c = 1.53$.
6. What is the value of c for which $P(Z \leq c) = .0162$? Answer: $c = -2.14$.

- We are usually interested in a characteristic of the population, like the mean μ.
- The mean is the balance point of the probability distribution, and equals the median when the distribution is symmetric.
- Another important feature of a distribution is the variance, σ^2, which is a measure of the spread of the values about the mean.
- The standard deviation, $\sigma = \sqrt{\sigma^2}$, is in the same units as X.
- If a distribution is symmetric and unimodal (i.e. looks approximately like a normal distribution), then
  - approximately 68% of the probability is between $\mu - \sigma$ and $\mu + \sigma$
  - approximately 95% of the probability is between $\mu - 2\sigma$ and $\mu + 2\sigma$
  - approximately 99.7% of the probability is between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Case I: estimation of a population mean

- Denote the values in a random sample as $X_1, \ldots, X_n$.
- We estimate the population mean μ using the sample mean

$$\bar{X} = \frac{\sum X_i}{n} = \frac{1}{n}(X_1 + \ldots + X_n)$$

- The **Law of Large Numbers** is a theoretical result which assures us that the sample mean $\bar{X}$ will be close to the population mean μ in large random samples.
- Before we have data, the sample mean is itself a random variable.
- Its probability distribution is called the **sampling distribution** - different samples give different values of the sample mean.
- The sample mean is an unbiased estimator because its sampling distribution is centered about the population mean μ.
- A measure of error of estimation of μ by $\bar{X}$ is the **standard error**, which is the standard deviation of the sampling distribution of $\bar{X}$.
- The standard error is much smaller than the standard deviation of the population, and is

$$\sigma_{\bar{X}} = \sigma/\sqrt{n}$$

- The larger the sample, the smaller the standard error, and the more accurate the estimate.
- We estimate the variance of the population using the sample variance

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum(X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\left(\sum X_i^2 - (\sum X_i)^2/n\right) \\ &= \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right) \end{aligned}$$

- We estimate the standard error of the mean by

$$s_{\bar{X}} = s/\sqrt{n}.$$

- If the populaton itself was normally distributed, then the sampling distribution of the mean has a normal distribution.
- A remarkable fact is that the sampling distribution of $\bar{X}$ is approximately normal (i.e. bell shaped) in large samples regardless of the shape of the probability distribution for the population. A sample size of 35 or larger is usually big enough so that the distribution of the sample mean is approximately normally distributed. This is the **Central Limit Theorem**, one of the most important results in statistical theory.

Case II - estimation of a proportion using binary data

- In clinical trials, we are often interested in whether a therapy is a success ($X = 1$) or a failure ($X = 0$).
- A model of the population is a Bernoulli random variable and distribution, which says $P(X = 1) = p$ and $P(X = 0) = 1 - p$.
- The quantity of interest is p, the probability of a success, which plays the role of the mean μ in this simple case.
- The sample mean (of the 0 or 1 variables) is just the sample proportion

$$\bar{X} = \hat{p} = \frac{1}{n}(X_1 + \ldots + X_n)$$

- The variance of the Bernoulli random variable is

$$\sigma^2 = p(1 - p)$$

  so we estimate the standard error of $\hat{p}$ using

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- Once again the standard error is inversely proportional to the square root of the sample size, and so gets smaller as the sample size gets large.