Chap6

Discrete: Number of outcomes can be counted

E.g. Roll of die

Continuous: Outcome occurs over a range

E.g. Exact time to finish exam

Residual: observed value-predicted value

Normal Model: continuous

Binomial Model: discrete

b(n,p,x): n->fixed number of trials, p->probability of success, x->number of success interested

The Binomial Model

- Discrete probability distribution
  - the outcome variable is discrete (counts)
- All 4 characteristics must be present:
  - A fixed number of trials.
  - Only two outcomes are possible at each trial, "success and failure". Ex. heads/tails, male/female.
  - The probability of success is the same at each trial.
  - The trials are independent.

μ=np ; Stdev=√(np(1-p))

Chap7

Sample entire population: census

Voluntary response samples are almost always biased, not representative

Non-response error occurs when people respond differ from people don't

Simple random Sampling(SRS): Randomly draw people from population without replacement

Bias

- A method is biased if it has a tendency to produce an untrue value.
- Sampling bias results from taking a sample that is not representative of the population.
  - Convenience sampling
  - Voluntary response sampling
- Measurement bias comes from asking questions that do not produce a true answer.
  - Confusing wording, misleading questions.

Population vs Sample

- Population is the collection of ALL data values.
- Population size is usually very large, often unknown, and usually impossible to obtain all values.
- Measures that come from the population are parameters.
- Sample is a subset of the population.
- Sample size (n) is the number of observations in a sample.
- Measures that come from the sample are statistics.

Accuracy and Precision

- The accuracy is measured in terms of the bias (taking a good sample).
  - If only basketball players are measured to estimate the proportion of Americans who are taller than 6 feet, then there is a bias for a larger proportion.
- The precision is measured by a number called the standard error (sample size).
  - If the sample size is only three, the estimate of the proportion of tall people using the sample is likely to be far from the proportion of tall people in the US. The sample size is small. The estimation method is not very precise.

Precision: spread; Accuracy: mean

Precision and bias are independent of population size as long as population size is at least 10 times larger than the sample

Bias and Standard Error

- Bias and standard error are easy to find for a sample proportion under certain conditions.
  - The sample must be randomly selected from the population of interest, either with or without replacement
  - If the sampling is without replacement, the population needs to be much larger than the sample size; at least 10 times bigger.
- Once these conditions are met, bias of $\hat{p}$ is 0 and the standard error is $SE = \sqrt{\frac{p(1-p)}{n}}$
- Note: In real life, we don't know the true value of the population proportion, p. This means we can't calculate the standard error, but we can come pretty close by using the sample proportion.

n: number of people in sample

The Central Limit Theorem for Sample Proportions

The Central Limit Theorem for Sample Proportions tells us that if some basic conditions are met, then the sampling distribution of the sample proportion is close to the Normal distribution.

- Random and Independent: The sample is collected randomly and the trials are independent of each other.
- Large Sample: The sample size is large enough that the sample expects at least 10 successes $np \geq 10$ and 10 failures. $n(1-p) \geq 10$
- Big Population: If the sample is collected without replacement, then the population size is at least 10 times larger than the sample size. $N \geq 10n$

CLT for Sample Proportion:

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \quad SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Interval $\hat{p} \pm \underbrace{z^* SE(\hat{p})}_{ME}$

Difference in population mean

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Random sampling is used in observational studies, allowing us to make inferences from the sample to the population
Random assignment is used in experiments, allowing us to possibly conclude causation

Chap8

Significance level
The probability of making the mistake of rejecting H_0 when H_0 is true(Type I error)
We never "accept" but only "fail to reject" H_0
If H_0 correct, p-value should be close to 0
Hypothesis test for one/two proportion

$$z = \frac{\text{observed value - null value}}{SE} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

The Test Statistic

- We are interested in how p_1 and p_2 differ, so our test statistic is based on the difference between our sample proportions from the two populations.
- The two-proportion z-test statistic is:

$$z = \frac{\text{estimator - null value}}{SE} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{where } \hat{p} = \frac{\text{number of successes in sample 1+ number of successes in sample 2}}{n_1 + n_2}$$

If p-value<a, reject H0, sufficient evidence that Ha is plausible
If p-value>a, fail to reject H0
Smaller p-value means easier to reject H0(less possible that H0 is true)

Chap9

Central Limit Theorem for population mean
Must assume that the population is normally distributed

Hypothesis Testing for 2 Sample Means

- Step 1: Hypothesize

  H_0: $\mu_1 = \mu_2$

  H_a: $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 \neq \mu_2$
- Step 2: Two-sample t-test and CLT Conditions:
  - Random Samples and Independent Observations
  - Independent Samples
  - Large Sample
- Step 3: Test statistic $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- Step 4: P-value
- Step 5: Interpret the results

Types of Mistakes

The first type of mistake is to reject the null hypothesis when it is true. This is also known as a **Type I Error**.

The second type of mistake is to fail to reject the null hypothesis when it is false. This is also known as a **Type II Error.**

Type I: false positive, H0 is positive
Type II: false negative, H0 is negative

Use z-test for proportion->categorical;
Use t-test for mean->numerical

Confidence Intervals for Means

We estimate the unknown σ (population standard deviation) by the known s (sample standard deviation) and calculate the standard error of $\bar{x}$.

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

The confidence interval for a mean is:

$$\bar{x} \pm t^*_{df} \frac{s}{\sqrt{n}}$$

where df = n - 1 and t* is found using the t-distribution.

The larger the df, the thinner the tails and distribution is more similar to N(0,1)