Biostatistics

Lecture 5

Estimation

Random Samples

- Now we want to infer the characteristics of a population from a representative sample
  - Usual strategy is through a random sample of some form
    - Simple random samples
    - Stratified random samples
    - Cluster sample
- Random number tables
  - Basic principal is that each digit has a equal probability of being selected as a random number

Randomized Clinical Trials

- Basic principal of randomized clinical trials is the random assignment of patients to treatments
  - Increases the probability that the treatment groups will be balanced for important baseline patient characteristics that might influence outcome
- Design features
  - Block randomization
  - Blinded assessments
    - Double vs. single
  - Endpoints / outcomes
  - SAEs

Sampling Distributions

- Definition
  - Distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, called the sampling distribution of that statistic

Sampling Distributions

- Construction
  - From a finite population of size N, randomly draw all possible samples of size n
  - Compute the statistic of interest for each sample
  - Investigate the distribution of that statistic
- Why??
  - Determine sampling distribution for a new statistic

Distribution of Sample Mean

- Possible the most important of all sampling distributions
  - Basis for much of statistical inference
- Sample Mean
  - Simple mean of the sample
    - $\overline{x}$
  - In theory, mean of sampling distribution has same mean as original population

Sampling Distribution of $\bar{x}$

- Variance
  - Variance of sampling distribution of $\bar{x}$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

  - Standard deviation of sampling distribution of $\bar{x}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  - Called the Standard Error

Sampling Distribution of $\bar{x}$

- Sampling from a normally distributed population
  - Distribution of $\bar{x}$ will be normal
  - Mean of the distribution of $\bar{x}$ will be equal to the mean of the original population
  - Variance of distribution of $\bar{x}$ will be equal to the standard error or the variance of the original population

Sampling from Nonnormally Distributed Populations

- Central limit theorem
  - Extremely important
  - Given a population of any nonnormal functional form with a mean 💻 and finite variance ☞^{2} , the sampling distribution of $\bar{x}$ computed from samples of sizes n from this population, will have mean 💻 and variance ☞^{2}/n and will be approximately normally distributed when n is large

Sampling from Nonnormally Distributed Populations

- Importance of central limit theorem is that we can sample from nonnormal distributions and still apply the standard normal theory procedures if sample is large enough
- Important for statistical inference

Implications of Central Limit Theorem

- Assured of approximately normal sampling distribution
  - Normally distributed population
  - Nonnormal population with large sample size
  - Population with unknown functional form with large sample size
- "Large" sample size
  - at least 30, but as many as practical if functional form unknown

Sampling without replacement

- When sampling from a finite population without replacement, the sampling distribution of $\bar{x}$ will have mean 💻 and variance $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$
  - Finite population correction
- When n is large, central limit theorem applies and sampling distribution of $\bar{x}$ will by approximately normal

Sampling Distributions Summary

- Normal population

$$\mu_{\bar{x}} = \mu$$

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

  - Sampling distribution of $\bar{x}$ is normal

- Nonnormal population
  - Same except for finite population correction

Interval Estimation or Confidence Intervals

- Having estimated a sample mean and standard deviation, how confident are we that the sample mean is close to the true mean
  - Construct a confidence interval with a specified level of 'confidence'
    - a 95% confidence interval means that the true mean will be contained in the confidence interval 95% of the time

$$Z = (\overline{X} - \mu) / (\sigma / \sqrt{n})$$

  - If σ is known or sample size large, then can use the standard normal distribution
  - Otherwise, use the t-distribution with n-1 degrees of freedom

Interval Estimation or Confidence Intervals

- Confidence interval for the mean of a normal distribution with unknown population variance

$$\overline{X} \pm t_{(n-1,1-\alpha/2)}\, s/\sqrt{n}$$

- Thus, for a 95% confidence interval for a sample with 20 observations,

$$\overline{X} \pm 2.093\, s/\sqrt{n}$$

  - 2.093 is the t(19,0.975) percentile point from Table 4

Interval Estimation or Confidence Intervals

- Confidence interval for the mean of a normal distribution with known population variance or large sample size (n > 200)

$$\overline{X} \pm Z_{1-\alpha/2}\, s/\sqrt{n}$$

- Thus, for a 95% confidence interval for a sample with 200 observations,

$$\overline{X} \pm 1.96\, s/\sqrt{n}$$

  - 1.96 is the z(.975) percentile point from Table 3

Confidence Interval for Variance

- Confidence interval for the variance of a normal distribution
  - Depends on the chi-square distribution, which is the distribution of squared normal variables
  - So, if underlying distribution is not normal, the estimation will be poor
- $(n-1)s^2 / \chi^2_{n-1,1-\alpha/2}$, $(n-1)s^2 / \chi^2_{n-1,\alpha/2}$
  - Not symmetric
- Rarely used in practice as such, but equality of variances are tested (using the F-distribution) in a number of procedures

CI for Difference between Two Population Means

- Two random samples are drawn from two populations
- Confidence interval for difference between two means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Use sample variance for sample means
- When sample size is < 200, use t distribution

CI for Difference between Two Population Means

- Variances equal
  - Can use pooled estimate of variance as on previous slide except that a pooled estimate of the sample variance is used
  - $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
- Variances unequal
  - Use unpooled estimate as on previous slide

CI for Population Proportions

- Same procedures as with means, except substitute p for $\bar{x}$ and p(1-p) for s^2
- For example,

$$\hat{p} \pm z_{(1-\alpha/2)} \sqrt{\hat{p}(1-\hat{p})/n}$$

- Is the 100(1-✂) percent confidence interval for p
- Similarly for difference between two population proportions
- Exact methods very cumbersome
  - Software available if necessary
  - Table 7a and b will give intervals

CI for Poisson Parameter

- Normal theory methods do not exist for the Poisson
- Table 8 presents intervals

One-sided Confidence Interval

- Same methods as before, but we use the z- or t-distribution percentile points at 1-α rather than 1- $\alpha/2$
- Thus, for the z-distribution, we use 1.645 instead of 1.96
- Otherwise, the other elements of the CI formula is the same