Review of Week 5 Biostatistics

For this handout, we consider a random sample of independent and identically distributed observations $(X_1, X_2, ..., X_n)$. The true mean of X_i in the population is μ and true variance is σ^2. The sample mean is $\bar{x} = (1/n)\sum_{i=1}^{n} X_i$. The sample variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$; and the sample standard deviation is $s = \sqrt{s^2}$.

- **Sampling Distributions.** In statistics, when we sample from a population, the individual observations in our sample are considered random variables. The mean of the observed observations $\bar{x}$, called the sample mean, is also a random variable. To understand this concept, think about the fact that we have only observed *one sample* out of the many different random samples that we could have observed.

- **Central Limit Theorem.** The CLT tells us about the behavior of the sample mean $\bar{x}$ in large sample sizes. If we take a random sample of size n, $(X_1, X_2, ..., X_n)$, from a population with true mean μ and standard deviation σ, the CLT says that, for large n,

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

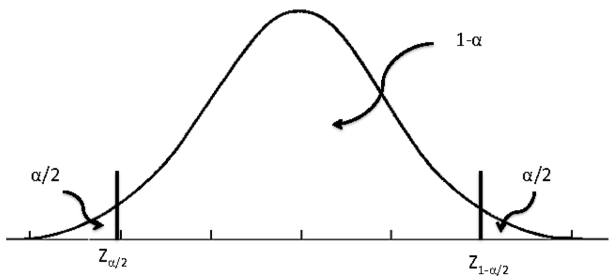
  - Using the Central Limit Theorem, we can assume that, in large samples, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$.
  - When the CLT applies, we can calculate the probability that $\bar{x}$ lies between a and b very easily (when σ is known), now that we know the sampling distribution of $\bar{x}$.

    Specifically, $P(a < \bar{x} < b) = P(\bar{x} < b) - P(\bar{x} < a) = P(Z < \frac{\bar{x}-b}{\sigma/\sqrt{n}}) - P(Z < \frac{\bar{x}-a}{\sigma/\sqrt{n}})$. To calculate these numbers, we can simply apply the `normal` function in Stata.

- **Predictive interval.** Predictive intervals are used to make predictions about random samples from a population, assuming the population parameters are known. Assume $X_i \sim N(\mu, \sigma^2)$.

  - A 95% predictive interval for X_i is $\mu \pm Z_{0.975}\sigma$.
  - A 90% predictive interval for X_i is $\mu \pm Z_{0.95}\sigma$.
  - A 95% predictive interval for $\bar{x}$ is $\mu \pm Z_{0.975}\sigma/\sqrt{n}$.
  - A 95% predictive interval for $\bar{x}$ is $\mu \pm Z_{0.95}\sigma/\sqrt{n}$.
  - For the 95% predictive interval, to calculate $Z_{0.975}$, use `di invnormal(0.975)`.
  - For the 95% predictive interval, to calculate $Z_{0.95}$, use `di invnormal(0.95)`.

- More generally, for a $1-\alpha$ confidence interval, to calculate $Z_{1-\alpha/2}$, use `di invnormal(1-`$\alpha/2$`)`. See the picture below, depicting a normal curve.



- **Confidence interval.** Confidence intervals are used to make inference about the population mean μ using a random sample from the population. Assume $X_i \sim N(\mu, \sigma^2)$, where μ is unknown.
  - When σ is known, a 95% confidence interval for μ is $\bar{x} \pm Z_{0.975}\sigma/\sqrt{n}$.
  - When σ is known, a 99% confidence interval for μ is $\bar{x} \pm Z_{0.995}\sigma/\sqrt{n}$.

  - When σ is unknown, a 95% confidence interval for μ is $\bar{x} \pm t_{0.975,(n-1)}s/\sqrt{n}$.
  - When σ is unknown, 99% confidence interval for μ is $\bar{x} \pm t_{0.995,(n-1)}s/\sqrt{n}$.

  - For the 95% confidence interval, to calculate $t_{0.975,(n-1)}$, use `di invttail(n-1, 0.025)` when σ is unknown.
  - For the 95% confidence interval, to calculate $t_{0.95,(n-1)}$, use `di invttail(n-1, 0.05)` when σ is unknown.
  - More generally, for a $1-\alpha$ confidence interval, to calculate $t_{1-\alpha/2,(n-1)}$, use `di invttail(n-1, `$\alpha/2$`)` when σ is unknown.

- **Confidence versus Predicive Intervals.**
  - Predictive intervals: Use the population distribution (known) to make predictions about a sample from the population (unknown). (This is the basis of probability.)
  - Confidence intervals: Use a random sample from the population (known) to make inference about the population distribution (unknown). (This is the basis of statistics!)

- **Hypothesis Testing.** Using the random sample, we want to test $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$.
  - **One-sample Z-test.** For large samples (CLT) OR when $X_i \sim N(\mu, \sigma^2)$ and σ is known, the test-statistic is:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

    Under H_0, $Z \sim N(0, 1)$.

- * For the one-sided test with alternative hypothesis $H_a : \mu > \mu_0$, we can calculate a p-value using the formula $p = P(Z > Z^*)$.
  * For the one-sided test with alternative hypothesis $H_a : \mu < \mu_0$, we can calculate a p-value using the formula $p = P(Z \leq Z^*)$.
  * For the two-sided test with alternative hypothesis $H_a : \mu \neq \mu_0$, we can calculate a p-value using the formula $p = 2 * P(Z < -|Z^*|)$.

- **One-sample t-test.** When $X_i \sim N(\mu, \sigma^2)$ and σ is unknown, the test-statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  Under H_0, $t \sim t_{n-1}$ (the test statistic follows a t-distribution with $n-1$ degrees of freedom.)