Stat 381: Applied Statistical Methods I **Spring 2015**

April 3-10

Instructor: Brian Powers

17.1 Statistical Inference and Classical Estimation

Definition 17.1. A **point estimate** of a population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$.

Example 17.2. The value $\bar{x}$ of $\bar{X}$ is a point estimate of population mean μ. $\hat{p}$ is a point estimate of the true proportion p for a bernoulli or binomial experiment.

Some properties make a "good" estimator. One is that on average the estimator is correct, this is called unbiasedness.

Definition 17.3. A statistic $\hat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if

$$E(\hat{\Theta}) = \theta.$$

Example 17.4. Show S^2 is an unbiased estimator of σ^2.

If $\hat{\Theta_1}$ and $\hat{\Theta_2}$ are to unbiased estimators for θ the "better" one is the one with the smaller variance. We refer to this as being more efficient.

Definition 17.5. If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the **most efficient estimator** of θ.

Example 17.6. Take a sample of $X_1, \ldots, X_n$ from a normal distribution with variance 25 and unknown mean. X_1, $\frac{X_1+X_2}{2}$ and $\bar{X}$ are all unbiased but $\bar{X}$ is the most efficient.

Another property desirable in an estimator is that of "consistency." Without being too technical, we say that an estimator is **consistent** if as the sample size increases, the estimator converges, in some sense, to the parameter θ.

17.2 Confidence Intervals for μ

Let us assume that σ is known and that the sample size is large enough for us to approximate the sampling distribution of $\bar{X}$ by a Normal distribution, ala the Central Limit Theorem. Suppose we want to create an interval which we are fairly confident contains the population mean. Typically we denote our desired confidence level as $100(1-\alpha)\%$. To achieve this, we want the upper and lower bounds of the interval to trap

$\alpha/2$ in each tail. We denote the z-values which do this as $-z_{\alpha/2}$ and $z_{\alpha/2}$. So we have

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= 1-\alpha \\ P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) &= 1-\alpha \\ P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha \\ P\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha \end{aligned}$$

Of course, this requires that σ, the population standard deviation is known. Furthermore, we should be careful how we interpret this probability. The population mean μ is fixed but unknown, we we cannot say "The probability that μ takes a value between this and that is $1-\alpha$." The random variable is $\bar{X}$. We must say "The probability that the interval from this to that contains μ is $1-\alpha$."

Example 17.7. Say we want to give a 92% confidence interval for the weight of frogs in a certain pond. We get a random sampling of 65 frogs and calculate the sample mean weight $\bar{x} = 22$ oz. Assuming we know the population distribution is symmetric with $\sigma = 5$ oz., the 92% confidence interval is as follows:
First a quick lookup or calculation gives us $z_{.04} = 1.751$. So the bounds of the confidence interval are

$$22 \pm (1.751)\frac{5}{\sqrt{65}} \approx 22 \pm 1.086.$$

We can be 92% confident that the interval $(20.914, 23.1086)$ contains the population mean.

17.3 Fixing the Margin of Error

The expression $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called the **margin of error** for a symmetric confidence interval as this. If we have control over the sample size and a specific margin of error as our target, we can find the minimum sample size needed to ensure our confidence level.

Say we want our margin of error to be no larger than ϵ. This is to say

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \epsilon.$$

By solving for n, we get

$$\left(\frac{z_{\alpha/2}\sigma}{\epsilon}\right)^2 \leq n.$$

We notice that in order to halve our margin of error, we must quadruple the sample size.

Example 17.8. Suppose we want the margin of error for the frog estimation to be .5 oz. How big must the sample size be?
We can just use this formula.

$$n \geq \left(\frac{1.751 \cdot 5}{.5}\right)^2 = 306.6$$

so we would need a sample size of 307 frogs to achieve this.

17.4 One-Sided Confidence Bounds

We may wish to simply find a lower bound on μ with a certain level of confidence. The procedure is similar, except we will use z_α instead of $z_{\alpha/2}$.

Example 17.9. Suppose the lifespan of transistors follows a normal distribution with unknown μ and $\sigma = 3$ years. A 25 random transistors are sampled and their average lifespan is 6.2 years. We want a lower bound on mean lifespan for which we are 95% confident.
For this we find $-z_{.05} = -1.645$. The lower bound on this confidence interval is

$$\bar{x} - z_{.05}\frac{\sigma}{\sqrt{n}} = 6.2 - 1.645\left(\frac{3}{\sqrt{25}}\right) = 5.213 \text{ years.}$$

17.5 When σ is Unknown

It is much more likely that the population standard deviation is unknown to us. Since we have collected a sample, though, we may naturally wish to use s, the sample standard deviation as a surrogate. How does this change the sampling distribution of $\bar{X}$? If an i.i.d. sample is drawn from a normal distribution,

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

will follow a Student's t-distribution with $n - 1$ degrees of freedom. We can now set up the confidence interval in a similar way. We will get

$$P\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

In practice, as long as the population distribution is approximately bell-shaped we can use this to get very good confidence intervals.

Most textbooks will argue that when the sample size is at least 30 and the population distribution is not too skewed, the Central Limit Theorem will ensure that the sampling distribution of $\bar{X}$ is approximately normal and we can forget about using a t-distribution. This is more justifiable as the sample size gets quite large, as "not too skewed" is not well-defined.

17.6 Standard Errors of Point Estimates

The standard error of a point estimate is the standard deviation of its sampling distribution. If σ is known,

$$\text{s.e.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

and when σ is unknown, the *estimated* standard error is

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}.$$

17.7 Estimating the Difference between Two population Means (Independent Samples)

From earlier, we know that if independent samples of sizes n_1 and n_2 are drawn from two populations with respective means μ_1, μ_2 and σ_1, σ_2, we can easily calculate the confidence interval. A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

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

If **variances are unknown but equal**, the standard error would be

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

And the best estimate for σ would be based on pooling the data from both samples,

$$S_{pooled}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

But rather than using the normal distribution, we have to utilize a t-distribution with $n - 2$ degrees of freedom.

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

When **variances are unknown and assumed unequal**, we have a complicated formula for v, the degrees of freedom for the t-distribution.

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]}$$

In practice, you should round down to the nearest integer.

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

17.8 Paired Observations

When there is a good reason to pair observations from two samples (such as before/after observations, without drug treatment/with drug treatment, husband/wife from couples), and we are interested in the difference in population means, we actually calculate the difference of each pair, and we treat the differences as a single population sample (where known/unknown variance is handled as before). Basically, the confidence interval is

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

where $\bar{d} = d_1 + d_2 + \cdots + d_n$ is the average of the paired differences, and s_d is the standard deviation of the paired differences. The t-distribution has $n - 1$ degrees of freedom.

17.9 Confidence Intervals for Population Proportion

Suppose we instead are looking at finding a confidence interval for a population proportion p. We begin by taking a sample proportion $\hat{p}$, which will serve as the point estimate. The sampling distribution of the statistic $\hat{P} = X/n$ can be calculated, where X follows Binom(n,p).

$$E(\hat{P}) = E(X/n) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$Var(\hat{P}) = Var(X/n) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}npq = \frac{pq}{n}$$

And the central limit theorem tells us that as $n \to \infty$,

$$\frac{\hat{P} - p}{\sqrt{pq/n}} \to N(0,1)$$

In practice, of course, the variance is not known - if it was we would already know p. As long as p is not close to 0 or 1, $\hat{p}$ may be used in its place, giving

$$\frac{\hat{P} - p}{\sqrt{\hat{p}\hat{q}/n}} \to N(0,1)$$

As before, the sampling distribution leads us to a $100(1-\alpha)\%$ confidence interval.

$$P(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{\hat{p}\hat{q}/n}} < z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < \hat{P} - p < z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = 1 - \alpha$$

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = 1 - \alpha$$

Example 17.10. A candidate's campaign staff conducts a random poll of 400 voters and finds that 250 of them plan to vote for her. Based on a 95% confidence interval, should she expect to win the election?

$\hat{p} = 250/400 = .625$, and $z_{.025} = 1.96$. Plugging into the formula, the bounds of the confidence interval will be

$$.625 \pm 1.96\sqrt{\frac{(.625)(.375)}{400}} = .625 \pm .047$$

17.10 Determining sample size for fixed margin of error

As with the population mean, we may have a target margin of error, that is to say some ϵ for which we need

$$z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq \epsilon$$

Solving for n we get

$$n \geq \frac{z_{\alpha/2}^2\hat{p}\hat{q}}{\epsilon^2}$$

But there is an obvious problem. If we have not yet determined the sample size, how could we have a sample proportion? There are two solutions. First we could do a small preliminary sample and use this $\hat{p}$ to determine the sample size for the full sample. You may argue that this estimate may be off and we should construct a confidence interval for it... seems quite convoluted. The other solution is to use a worst case scenario for $\hat{p}\hat{q}$. By this we mean that

$$\hat{p}\hat{q} = \hat{p}(1-\hat{p}) = \hat{p} - \hat{p}^2 \leq \frac{1}{4}$$

So plugging this value into the formula we get

$$n \geq \frac{z^2_{\alpha/2}}{4\epsilon^2}$$

Example 17.11. Suppose another candidate wants to conduct a poll and estimate the proportion of voters who support him with a margin of error of 1 percentage point at 99% confidence. The formula above says that he will need to survey at least

$$\frac{2.576^2}{4(.01)^2} = 16589.44$$

So at least 16,590 people!

17.11 Difference of two population proportions

If samples of sizes n_1 and n_2 are drawn from two populations giving estimates $\hat{p}_1$ and $\hat{p}_2$ of the two respective population proportions, and we want to give an interval estimate for the difference $p_1 - p_2$, we proceed similarly as with the 2-independent sample interval for μ. The standard error of $(\hat{p}_1 - \hat{p}_2)$ is

$$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

So a $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

17.12 Estimating the Variance

As you may assume, we use S^2, the sample variance as a point estimate of σ^2 usually, since S^2 is unbiased, consistent and is efficient. But recall that the statistic

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

follows a chi-squared distribution with $n-1$ degrees of freedom, when the samples are drawn independently from a normal population (and this is still approximately true when the population is not normal. We can construct the form of the confidence interval as we did for μ. We start with:

$$P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$$

Divide each side by $(n-1)S^2$ and taking a reciprocal we get

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

Where $\chi^2_{\alpha/2}$ leaves a probability of $\alpha/2$ to the right in a Chi-squared distribution with $v = n-1$ degrees of freedom. You can express the confidence interval for σ by taking the square root of the upper and lower bounds.

Example 17.12. We sample 10 packages of grass seed, measuring the weights. We get

$$46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, \text{and} 46.0$$

Find a 95% confidence interval for σ^2