QUAN 102 Statistics for Business

Tutorial **7**

Natt Krongkajonsook

7.1 Estimating The Mean from Small Samples

- Recall that if σ is known, the confidence interval for $\mu = \bar{X} \pm E = \bar{X} \pm Z_{\alpha/2}\sigma_{\bar{X}} = \bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Also recall that when σ is unknown, but n is sufficiently large ($n \geq 30$), we can replace σ by s without having to change any results.

- When σ is unknown, and $n < 30$, CLT no longer applies and therefore the mean is no longer normal. In this case, the mean has a Student's t distribution with ν degrees of freedom, provided that the population is normal. (*Here the required assumption is that the population is normal, not the mean. If the mean is normal, we can use the previous formula to calculate the confidence interval.*)

- The t distribution is very similar to the Z distribution.
  - It is symmetric about its mean.
  - It has a mean of zero.
  - It has a standard deviation and variance greater than 1.
  - There are actually many t distributions, one for each degree of freedom (ν). For a single mean, $\nu = n - 1$.
  - As n increases (or ν increases), the t distribution approaches the Z distribution. The Z distribution is actually the t distribution with ∞ degrees of freedom.
  - It is bell shaped just like the Z distribution, but with lower peak and fatter tails.
  - The t-scores can be negative or positive, but probabilities (areas under the curve) are always positive.

- If σ is unknown and $n < 30$, the confidence interval for $\mu = \bar{X} \pm E = \bar{X} \pm t_{\alpha/2}\sigma_{\bar{X}} = \bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$ provided that the population is normal. Here the area between 0 and the t-score is simply one-half of α.

- A table of t distributions is available, but presented differently from the standard normal table. In particular, the area to the right of the t score can be looked up across the top. ν of the distribution can be looked up along the left side. The value in the intersection of the row and the column is the t score that corresponds to the given area. To find $t_{\alpha/2}$, for example, we look for the column that corresponds to *one-tail probability value* $= .5 - \alpha/2$, and the row that corresponds to the required ν, and then simply read the t score of that particular row and column.

- Exercise 8.12: $\bar{X} = 10.83$, $s = 2.135$, $n = 10$, $\nu = 10 - 1 = 9$, $\alpha = .90$, and $t_{.90/2}$ from the t distribution with 9 degrees of freedom $= 1.833$ are all you need to answer this exercise. Do not forget to assume that the population is normal.

7.2 Hypothesis Testing

- Recall that population parameters (μ, σ, p) are NOT observable, but sample statistics $(\bar{X}, s, \hat{p})$ are.
- When a claim is made about the population parameter, we can check whether that claim is correct by comparing the observed value and the claimed value.
- A comparison between the observed value and the claimed value can be made by calculating the "test statistic" which equals (*the observed value - the claimed value*)/(*standard deviation of the observed value*). If the claim is likely to be true, the observed value and the claimed value should not be very different from each other. As a result, the test statistic should be close to zero.
- We will compare the test statistic to some threshold value (known as the "critical value"). If the size of the test statistic is larger than the size of the critical value, we conclude that the difference between the observed value and the claimed value is too large, and therefore the claim should be rejected.
- We can also use p-values (short for probability values) to make a decision to reject or not to reject the claim. We can think of the p-value as the probability that a claim is "TRUE". Hence, if the p-value of the claim is "too small", we will reject that claim. How small is "too small"? If the test is carried out at $\beta\%$ level of significance, we reject the claim if the p-value is less than $\beta\%$.
- Another way of testing the claim is to use confidence intervals. Recall that the confidence interval is the interval that is likely to contain the true parameter. Thus, if the claimed value does not lie within the confidence interval, the claim should be rejected.

7.3 Type of Tests

- The first step to performing hypothesis testing is to write down the null hypothesis (H_0) and the alternative hypothesis (H_1).
- If the given claim contains equality, or a statement of no change from the given or accepted condition, then it is H_0, otherwise, if it represents change, it is H_1. Thus H_0 always involves "$=$", whilst H_1 involves "$<$", "$>$", or "$\neq$".
- The type of test is determined by H_1.
  - If H_1 involves $<$, then the the test is lower tailed. The lower tailed test is usually carried out when the observed value is less than the claimed value.
  - If H_1 involves $>$, then the the test is upper tailed. The upper tailed test is usually carried out when the observed value is more than the claimed value.
  - If H_1 involves $\neq$, then the the test is two tailed. The two tailed test is usually carried out when we suspect that the true parameter does not equal the claimed value.

7.4 Testing The Mean

- The formula for the test statistic in this case $= \frac{\bar{X}-\mu^{H_0}}{\sigma_{\bar{X}}}$, where μ^{H_0} is the value of μ from H_0 (the claim), and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
- The distribution of the test statistic will depend on different situations. The following summarizes three situations that can arise.
  - If σ is known, then $\bar{X}$ is normal. The test statistic therefore has a standard normal distribution. If however σ is unknown, but $n \geq 30$, we can replace σ by s, and the test statistic still has the standard normal distribution.
  - If σ is unknown, and $n < 30$, we still replace σ by s, but now the test statistic has the t distribution with $\nu = n - 1$ degrees of freedom, provided that the population is normal.
  - If σ is unknown, $n < 30$, and the population is not normal, we need to test the population median ($\mathcal{M}$) instead. In this case, we need to assume that the population is symmetric, so that $\mu = \mathcal{M}$ and we therefore can draw the same conclusion for μ. Testing the median is discussed in the next section.
- We need to know the distribution of the test statistic to calculate the critical value or the p value, so that the decision to reject or not to reject the claim can be made.
  - For the lower tailed test carried out at $\beta\%$ level of significance, the critical value is the value that cuts off the lower tail of the distribution of the test statistic at $\beta\%$. The p value is the area to the left of the test statistic.
  - For the upper tailed test carried out at $\beta\%$ level of significance, the critical value is the value that cuts off the upper tail of the distribution of the test statistic at $\beta\%$. The p value is the area to the right of the test statistic.
  - For the two tailed test carried out at $\beta\%$ level of significance, critical values are values that cut off both tails of the distribution of the test statistic at $\frac{\beta}{2}\%$. In this case, we can calculate the p value by looking up the area in one tail and double it.
- Exercise 8.5: $n = 38 \geq 30$, $\bar{X} = 3.62$, $s = .78$
  - $H_0 : \mu = 3.94$ against $H_1 : \mu < 3.94$ is being carried out.
  - The test statistic $= \frac{\bar{X}-\mu^{H_0}}{\sigma_{\bar{X}}} = \frac{3.62-3.94}{.78/\sqrt{38}} = -2.52$ has the standard normal distribution.
  - The critical value of this lower tailed test is -2.33 at 1% level of significance. $|\text{test statistic}| > |\text{critical value}|$, therefore the difference between the observed value and the claim value is too large at 1% level of significance, we therefore must reject H_0.
  - The p value is the area (under the Z curve) to the right of -2.52, i.e. p value $= P(Z < -2.52) = .5 - P(-2.52 < Z < 0) = .5 - P(0 < Z < 2.52) = .5 - .4941 = .0059$. The p value is less than 5% meaning that the probability that H_0 is true is too small, we therefore must reject H_0.

- 99% confidence interval for $\mu = \bar{X} \pm z_{\alpha/2}\sigma_{\bar{X}} = 3.62 \pm 2.576 \times .78/\sqrt{38} = (3.29, 3.95)$. The claimed value lies within the interval, we therefore cannot reject H_0. *The conclusion here is different from that of the* 1% *lower tailed test because using* 99% *confidence interval to do hypothesis testing is equivalent to using the* 1% *two tailed test.*

- Exercise 8.18: $n = 15 < 30$, $\bar{X} = 10.2$, $s = \sqrt{5.29} = 2.3$, $\nu = n - 1 = 15 - 1 = 14$
  - $H_0 : \mu = 9.4$ against $H_1 : \mu \neq 9.4$ is being carried out.
  - The test statistic $= \frac{\bar{X} - \mu^{H_0}}{\sigma_{\bar{X}}} = \frac{10.2 - 9.4}{2.3/\sqrt{15}} = 1.347$ has the t distribution with 14 degrees of freedom, assuming that the population is normal.
  - Critical values of this two tailed test are -1.761 and 1.761 at 5% level of significance. |test statistic| < |critical values|, therefore the difference between the observed value and the claim value is NOT too large at 5% level of significance, we therefore cannot reject H_0.
  - The p value cannot be computed in this case because, unlike the Z table, the t table does not provide areas under the t curve.
  - We can use the 95% confidence interval to do the test and still get the same conclusion.

7.5 Testing The Median

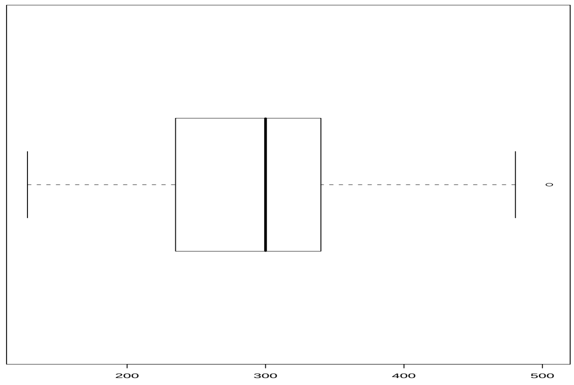
- Recall that $\mathcal{M}$ locates the center of the population. Almost any population is very large, and therefore there will be many observations that equal $\mathcal{M}$.

- When a claim is made about $\mathcal{M}$, which is unobservable, we collect the data to see whether it shows more observations that are above the claimed value. If this is the case, then $\mathcal{M}$ is likely to be greater than the claimed value. The claim will be reject if there are "too many" observations that are above the claimed value. *(The opposite is also true.)* The only question is how many is "too many"? Normally, we need to calculate the test statistic and then compare it to some critical value. Or equivalently, we can calculate the p value and see if it is too small for the claim to be true. The latter approach will be a lot easier in this case.

- The test statistic is X^+, which is the number of observations that are above the claimed value. $X^+ \sim Bin(n, p)$, where n is the number of observations in the data excluding ones that equal the claimed value, and $p = .5$ is the probability that an observation will be above $\mathcal{M}$.

- Similarly to the mean case, the p value will depend on the type of tests carried out:
  - If the data shows more observations that are above the claimed median, then we should test $H_0 : \mathcal{M} = \mathcal{M}^{H_0}$ against $H_1 : \mathcal{M} > \mathcal{M}^{H_0}$, where $\mathcal{M}^{H_0}$ is the claimed median. The p value in this case will be all binomial probabilities in RHS of the test statistic, i.e. p value $= P\left(Bin(n, p) \geq X^+\right)$.
  - If the data shows more observations that are below the claimed median, then we should test $H_0 : \mathcal{M} = \mathcal{M}^{H_0}$ against $H_1 : \mathcal{M} < \mathcal{M}^{H_0}$. The p value in this case will be all binomial probabilities in LHS of the test statistic, i.e. p value $= P\left(Bin(n, p) \leq X^+\right)$.

- Exercise 8.25



  - The boxplot shows that the data is asymmetric with an outlier. Testing for the mean is clearly not appropriate. We therefore must focus on the median instead.
  - The data gives us 13 below the claimed median ($\mathcal{M} = 413$) and 2 above it, so $H_0 : \mathcal{M} = 413$ should be tested against $H_1 : \mathcal{M} < 413$.
  - The test statistic is $X^+ = 2$.
  - The p value is $P\left(Bin(15, .5) \leq 2\right) = .057$. The test is carried out at 5% level of significance, we therefore cannot reject H_0 because the p value $< 5\%$.

- Now use everything you have learnt so far to answer Exercise 8.21.

- Testing the claim about $\mathcal{M}$ can sometimes be done using the two tailed test.
  - In this case, $H_0 : \mathcal{M} = \mathcal{M}^{H_0}$ should be tested against $H_1 : \mathcal{M} \neq \mathcal{M}^{H_0}$.
  - The test statistic is $X = \min(X^+, X^-)$, where X^- = the number of observations that are below $\mathcal{M}^{H_0}$. $X \sim Bin(n, p)$.
  - The p value can be found by doubling all binomial probabilities in LHS of the test statistic, i.e. p value $= 2 \times P\left(Bin(n, p) \leq X\right)$.