Chapter 15 : Tests of Significance

Note : Here the 4-th and 5-th editions of the text have different chapters, but the material is the same.

We have used confidence intervals to *estimate* the values of μ that are consistent with the data. In other words, these are the possible values of the population mean μ that are reasonable (consistent) with the observed data.

In many problems or studies there is a specific claim (or *hypothesis*) about the population or a parameter of the population. *IF* that hypothesis or claim *were* true, we can then measure in a probability sense, how unusual or rare is the observed data. If the observed data is *TOO RARE* then it suggests (perhaps strongly suggests or indicates) that the *hypothesis* or claim is not true.

This is analogous to the legal trial system where an accused is assumed innocent of a crime until *proven beyond a reasonable doubt* that innocent is *too inconsistent* with the observed data, or equivalently the observed data is *too unusual IF the accused is INNOCENT*.

In our statistical problem we need to measure *how rare*. For this we will use the term *level of significance* or *level of statistical significance*, which is defined a little more explicitly later.

We study this problem initially in the context of a random sample from a normal distribution, and only for the population mean parameter μ.

A simple prototype example to keep in mind (although this is not a normal population example) is a coin tossing game. You are playing a gambling game based on coin tosses with someone. They claim it is a fair coin, that is the coin has probability of heads coming up equals the probability of tails coming up, so both of these are $\frac{1}{2}$. Based on an experiment of tossing this coin some number n of times, we then examine if the observed number of heads (or sample proportion of heads) is consistent with this hypothesis (called null hypothesis). Our goal is to exam if this hypothesis is "true" or if an alternative

hypothesis is "true." Here a natural alternative hypothesis is that the coin is unfair against us since we would be happy if the coin were fair or biased in our favour, and be unhappy if the coin were biased against us and favoured our opponent.

Stating Hypotheses

The hypotheses are statements about the parameters for the population. They are NOT statements about the random variables that we observe.

Null and Alternative Hypotheses

The statement being tested is called the Null Hypothesis.

The alternative hypothesis is what we are trying to prove.

Hypothesis is a singular noun, and hypotheses is the plural of the word hypothesis.

In the coin tossing example as given about the hypotheses state something about the parameter p, the probability that the coin comes up heads. The null hypothesis is the statement that $p = \frac{1}{2}$ and the alternative hypothesis is $p > \frac{1}{2}$. This is referred to as a one sided alternative hypothesis.

Perhaps you are *gaming commissioner* and you are only interested in determining if the gambling game is fair. In this we consider the null hypothesis $H_0 : p = \frac{1}{2}$ versus the alternative hypothesis $H_a : p \neq \frac{1}{2}$. This alternative is called a two sided alternative.

In our test scores example I might claim that the population mean score is $\mu = 23$. As a student you might be interested in deciding if this is true, or if in fact the real population mean is smaller than 23, that is $\mu < 23$. In this case the null hypothesis is $H_0 : \mu = 23$ versus the alternative hypothesis $H_a : \mu < 23$.

If we take a random sample of size $n = 10$ and observe $\bar{x} = 22.3$, the observed value (statistic) is a number and not the value of the parameter μ. We do not make a hypothesis $\bar{x} = 22.3$. Hypothesis are only statements about the population model (population parameter in our case).

Since we wish to make our decision, that is whether the null hypothesis or the alternative hypothesis is most reasonable, based on observed data, we use a *test statistic* to help in making our decision.

Test statistic

- a test statistic is a statistic (something that can calculated from the observed data) that compares value of the (null) hypothesized value of the parameter with its estimated value based on the observed data
- large values of the statistic indicate estimate is far from value of the parameter as specified by the null hypothesis H_0.

As in our legal analogy the set up to assume the null hypothesis H_0 is true and then measure if the data is too rare when this is the case (too rare IF the null hypothesis H_0 were true).

Take a sample of size n from a Normal population with mean μ and standard deviation σ, where σ is a known value, for example 4.04.

Test the null hypothesis $H_0 : \mu = 21$.

$\bar{x}$ is a statistic that we used to estimate μ. It has the property (see Chapter 11) that

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

In particular IF THE NULL HYPOTHESIS H_0 were true then

$$\frac{\bar{x} - 21}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

More generally if the null hypothesis is

$$H_0 : \mu = \mu_0$$

for a specified value of μ_0 (for example 21) then IF H_0 were true

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

We use

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

as the test *test statistic*.

From the grades we take a sample of size $n = 15$ and observe the sample

$$26, 24, 25, 24, 29, 21, 20, 15, 29, 26, 25, 27, 21, 22, 29$$

This gives $\bar{x} = 24.2$. The observed value of the test statistic is

$$z = \frac{24.2 - 21}{\frac{4.04}{\sqrt{15}}} = 3.07$$

Suppose a student claims the average population score is $\mu = 21$ and we wish to examine this assumption versus the alternative that $\mu > 21$. We take $H_0 : \mu = 21$ and $H_a : \mu > 21$.

The evidence would favour the alternative when z is large and positive, since this would be the same as $\bar{x}$ much greater that $\mu_0 = 21$. In this case we would *reject* H_0 in favour of H_a. How do we decide what is large? It will be if

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z^*$$

where we still have to decide what is the *critical value* z^*.

If instead we consider $H_0 : \mu = 21$ and the alternative $H_a : \mu \neq 21$, the either z much smaller that 0 or much larger than 0 would both favour the alternative, as they correspond to $\bar{x}$ being far from 21. In this case of the two sided alternative we would reject H_0 in favour of the alternative if

$$|Z| = \left| \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| > z^*$$

or equivalently if

$$Z < -z^* \text{ or } Z > z^*$$

Of course we still need to decide what is the *critical value* z^* (different value than in the one sided alternative case above).

Instead of trying to think about finding the critical value z^* let us think about this idea in a slightly different way.

p-values

The probability, assuming that the null hypothesis is true, that the test statistic random variable is more extreme than the observed value of the test statistic for this sample is called the p-value of the test.

The smaller the p-value the more *rare* is the observed data or sample IF the null hypothesis WERE TRUE. A large p-value indicates the test statistic and hence the data is not so extreme IF the null hypothesis WERE TRUE.

The p-value plays the role we wanted in measuring how rare the data is if the null were true, the same as our legal analogy of innocent (null is true) until the evidence is that it is not true and hence the alternative hypothesis is true.

Notice the p-value is calculated in a form that depends on whether the alternative is a one sided or two sided alternative.

Test scores example continued

We observe the test statistic is $z_{obs} = 3.07$. The subscript obs is to indicate that this is the observed value of z with this data. The p-value is then

$$\begin{aligned} P(Z > z_{obs}) &= P(Z > 3.07) \\ &= 1 - P(Z \leq 3.07) \\ &= 0.001 \end{aligned}$$

This says that IF the null hypothesis $H_0 : \mu = 21$ WERE true then the chance of observing the test statistics score at least as extreme (in this case greater than equal to) $z_{obs} = 3.07$ is 0.001, that is only 1 in a thousand times for a simple random sample of $n = 15$ test scores. Since this p-value is extremely small it is quite *rare* (quite unbelievable) that H_0 is true and that it is more reasonable to think that $H_a : \mu > 21$ is in fact true.

Our conclusion is that the data indicates that the null hypothesis $H_0 : \mu = 21$ is not true and that the alternative hypothesis $H_a : \mu > 21$ is true.

The area under the sampling distribution curve for the test statistic is shown in Figure 1. Figure 2 shows the same plot but the upper tail is blown up for ease of viewing.

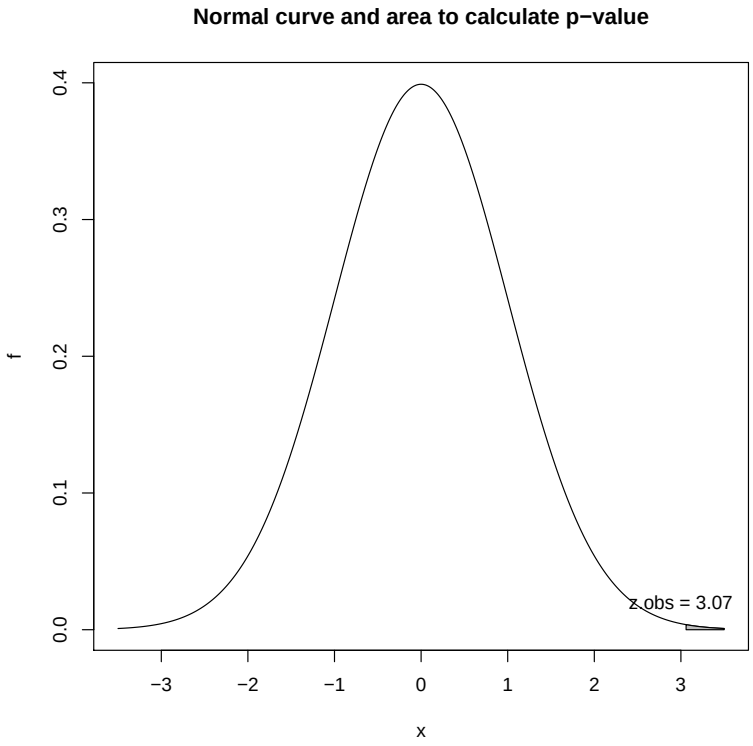

Figure 1: P-value for $H_0 : \mu = 21$ versus $H_a : \mu > 21$

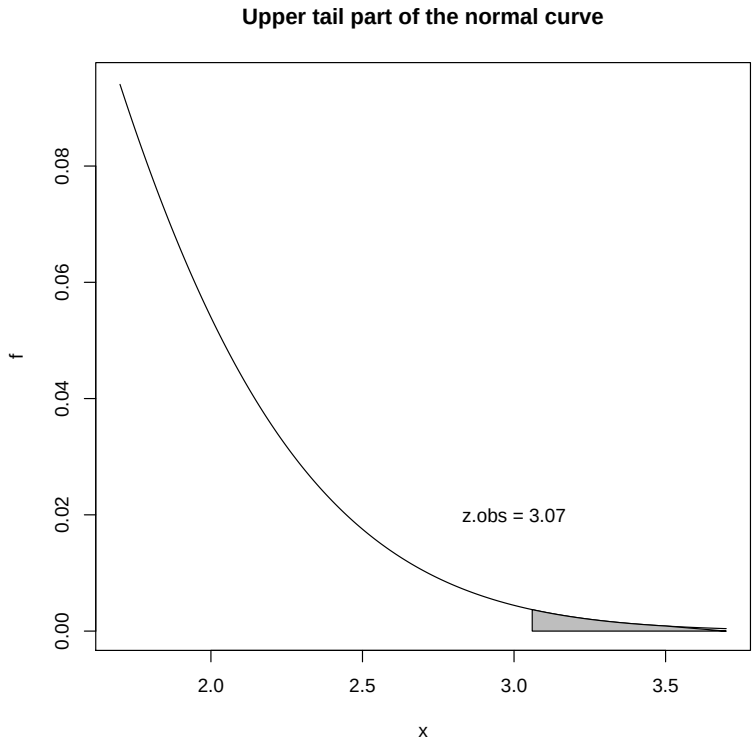

Figure 2: P-value for $H_0 : \mu = 21$ versus $H_a : \mu > 21$, Upper part of Normal Curve

CAUTION

The p-value does not calculate the probability that the null hypothesis is true. It only calculates the chance or probability that a more extreme value of the test statistic is observed IF the Null Hypotheses WERE true.

What constitutes a small p-value? Clearly 1 in a thousand is small. If 1 in 10 small? No, since there is a fairly good chance that such an event happens just by chance. Would you like to be convicted with a 1 in 10 chance if you were innocent? Typically there is a line of about 0.05 as a small p-value.

Statistical Significance

We specific a value of α (Greek letter alpha), say 0.05.

If the test statistic were to fall into the region that α chance of the test statistic being this rare, then we declare the test result significant and say there is evidence, at significance level α against the null hypothesis in favour of the alternative. This means that we set a critical value z^* so that

- in the case of $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0 : \ P(|Z| > z^*) = \alpha$
- in the case of $H_0 : \mu = \mu_0$ versus $H_a : \mu < \mu_0 : \ P(Z < z^*) = \alpha$
- in the case of $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0 : \ P(Z > z^*) = \alpha$

Using this critical value then we reject H_0 in favour of H_a if we observe (corresponding to the three alternative types above)

- $|z_{obs}| > z^*$
- $z_{obs} < z^*$
- $z_{obs} > z^*$

Note only one of these 3 alternative types is used for a given statistical hypothesis testing problem, NOT ALL THREE.

The above formulation is equivalent to :

If the p-value is $\leq \alpha$ then the data is statistically significant at level α.

Tests for a Population Mean

- Statist the practical question and phrase it in the form of a parameter for a statistical model.

  This may not always apply and so these methods do not apply in such cases.

- State the null hypothesis and alternative hypothesis
- Carry out the hypothesis test in three stages

  1. Check the conditions for the statistical test you plan to use
  2. calculate the test statistic
  3. find the p-value

- State your conclusions. This typically done in the context of the practical question.

There is a relationship between confidence intervals and the two sided alternative tests of hypotheses considered here.

A level α test of $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ rejects null hypothesis if and only if the parameter value μ_0 falls outside the $100(1-\alpha)\%$ confidence interval for μ.

We use the same random sample of the test 1 grades. The 95% confidence interval for μ is

$$\begin{aligned} \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= \bar{x} \pm 1.96 \frac{4.04}{\sqrt{15}} \\ &= 24.2 \pm 2.045 \\ &= [22.16, 26.24] \end{aligned}$$

Notice that with $\alpha = .05$, we have $100 * (1-\alpha) = 100 * 0.05 = 95$ percent.

On the other hand if wished to test $H_0 : \mu = 21$ versus the TWO SIDED alternative $H_a : \mu \neq 21$ at significance level $\alpha = .05$, then we know that we would reject this H_0 versus the two sided alternative since 21 is not in the interval $[22.16, 26.24]$.

Similarly if we wished to test $H_0 : \mu = 25$ versus the TWO SIDED alternative $H_a : \mu \neq 25$ at significance level $\alpha = .05$, then we know that we would not reject (that is accept) this H_0 versus the two sided alternative since 25 is in the interval $[22.16, 26.24]$.

We would have to redo the calculations to determine if we would reject $H_0 : \mu = 21$ versus the TWO SIDED alternative $H_a : \mu \neq 21$ at significance level $\alpha = .01$. For this we would need to use a 99% confidence interval, or find the appropriate critical value. If we have already calculated the 99% confidence interval then we obtain the result, that is do we accept or reject H_0, with no further calculations.