Math 321 – Chapter 6 – Hypothesis Testing
(draft version 2019/04/24-16:38:39)

Contents

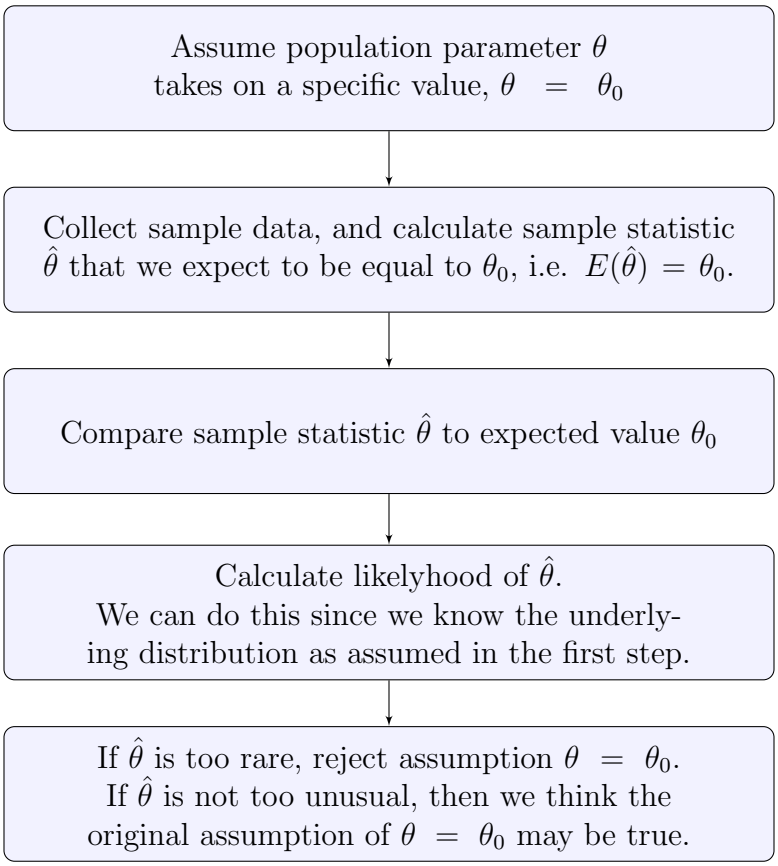

Figure 1: Explanatory flow of a hypothesis test.

1 Introduction

Now that you are able to construct confidence intervals, the next step is to understand hypothesis testing. There is a close relationship between the two statistical concepts.

Consider this situation: You are confronted with a particular claim about a statistical population, e.g. told the mean breaking strength of a particular construction material is 5,000 lb, however, you gather data and find that the mean breaking strength in your sample was only 4,900 lb. So you might begin to question the original claim as potentially being untrue. We are in the world of randomness though, and maybe 4,900 lb breaking strength is not all that unusual; maybe it's not too far a perturbation from the expectation of 5,000lb. Hypothesis testing is a way of determining if your data is sufficiently unusual to warrant for rejecting the initial claim of a mean of 5,000 lb.

The general process for a hypothesis test:

This will depend on our knowledge of sampling distributions.

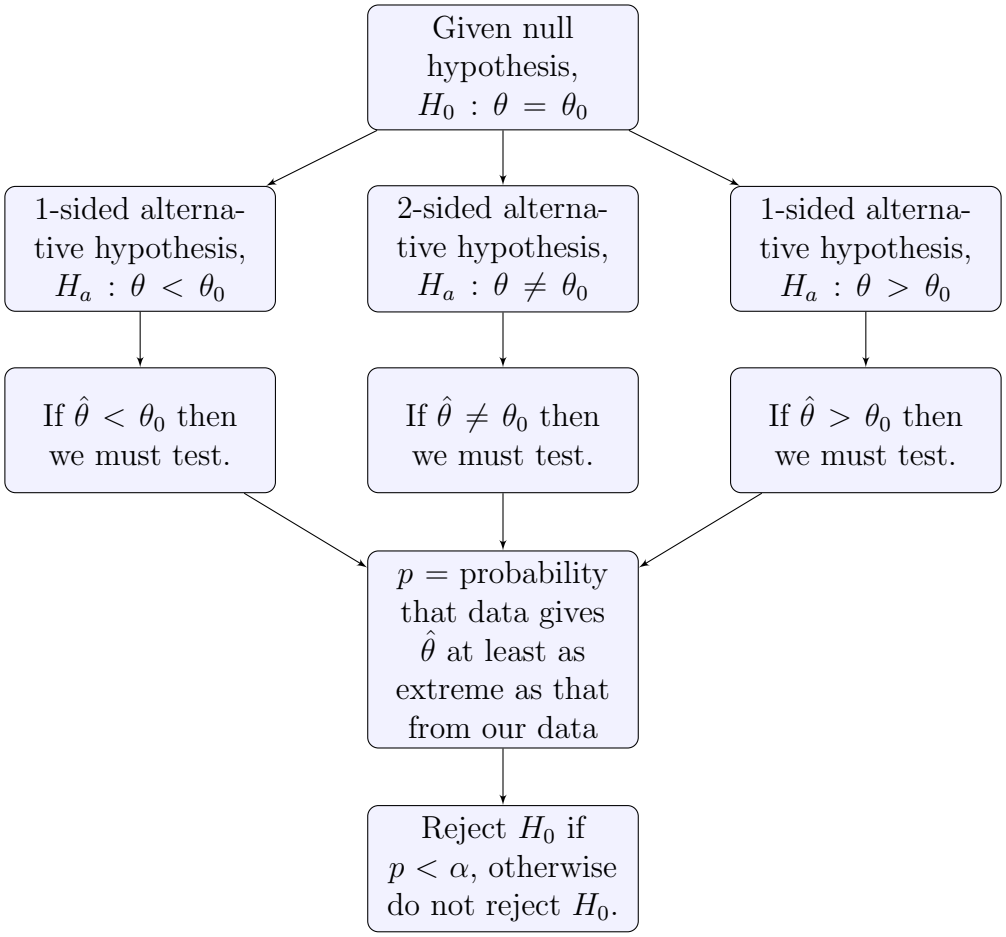

Figure 2: Technical flow of a hypothesis test.

1.1 A simple example hypothesis test

If I hand you a coin and claim that it is a fair coin, and you look at it and see that it has heads on both sides, would you believe me? No! You have collected data that leads you to believe my original claim is false. In fact you have collected all possible data effectively and are absolutely certain I am lying!

Let's say you don't look at the coin, but flip it once and get heads. Would you begin to question my claim? No, getting heads is typical. It has probability 50% under the assumption that the coin is fair.

What if you flip it twice, and, of course, get two heads in a row? Would you start to question my claim yet? Probalby not, since under the assumption of fairness, our data has probability 25%.

Do you see where this is going yet? How many heads in a row would you need to see in order to start questioning whether or not the coin is fair?

Under the assumption of fairness, getting n heads in a row has probability $\frac{1}{2^n}$.

We need to choose for ourselves, *how rare is too rare*. This is our chosen *level of statistical significance*. Let's say you will question fairness of the coin after 5 head sin a row. That has probability $1/2^5 = 3.125\%$. That is indeed a fairly low probability and is an unlikely outcome. So we have set 3.125% as our required significance level.

Maybe you don't think that is rare enough and you require 10 heads in a row. The you are choosing $1/2^{10} \approx 0.0977\%$ as your significance level.

A more mathematical explanation Let X_i be the outcome of the i^{th} coinflip. $X_i = 1$ for heads and $X_i = 0$ for tails. Each X_i is Bernoulli with parameter p, the probability of success.

We flip the coin n times and find that all flips are heads. Our sample mean of n Bernoulli trials is $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. It happens that the sample mean is also $\hat{p}_n$ which is our point estimate for p the true probability of heads. For our double-sided heads coin, we will always get $\hat{p} = 1$. Out of n coin flips, $P(\hat{p}_n = 1) = \frac{1}{2^n}$. This becomes exceedingly rare as we increase the number of flips, n. Let α be our chosen level of significance. If $\frac{1}{2^n} < \alpha$ then we reject the claim of fairness.

Eventually everyone now matter how demanding they are of rare-ness (no matter how small they set α will become satisfied that the coin is in fact not fair. Even further, I think we would all be quickly convinced that besides not being fair, clearly the coin is highly likely to be double-sided heads!

If we demand a very rare outcome, say $\alpha = 10^{-12}$ (1 part in a trillion!), then that would require $n \geq 28$ coin flips to get $\frac{1}{2^n} < \alpha$.

One final consideration is that if we are really just interested in whether the coin is precisely fair or not as opposed to whether it is double-sided heads vs fair, then that might change our decision making process. Rather than going so extreme as to require10 or more heads in a row to question the fairness of the coin, we may be satisfied with 5 heads in a row, since

that is already extremely rare. We don't need to try and confirm that it is double-sided. It could simply be biased with, say, 75% chance of heads.

If the true probability of heads is 75%, then the probability of 5 in a row is $(0.75)^5 \approx 23.7\%$. So for such a biased coin, 5 head sin a row is not all that unusual. One way to look at this is that 5 heads in a row is not yet enough to confirm that the coin is double-sided, but it is probably enough to begin to believe it is not fair.

Summary of this coin example

1. We are told a certain probability of heads.
2. We flip the coin a few times to get data. Count the number of heads.
3. Calculate the probability of getting this number of heads (or more†), under the assumed probability of heads.
4. If our data seems sufficiently rare, reject the initial claimed probability of heads.

† As we will see later, we are not interested in the probability of our specific outcome, but will be interested in the probability of any outcome that is at least as extreme as ours. What this means will depend on the context.

2 Terminology and methodology

The *null hypothesis* is a claim about a population parameter θ.

The *test statistic* $\hat{\theta}$ is calculated from the assumed null hypothesis and sample data.

The *significance level* is α, the probability that is our chosen level of "rareness." The most common value is $\alpha = 5\%$.

The *p-value* is the probability of having data at least as extreme or rare as ours.

The *rejection region* is a region where are test statistic must fall in order to have $p < \alpha$. Usually it will be of the form θ^* such that $p < \alpha$ when $\hat{\theta} < \theta^*$. There are several variations.

We reject the null hypothesis if $p < \alpha$ (or equivalently if $\hat{\theta}$ falls int eh rejection region).

We do not reject the null hypothesis if $p \geq \alpha$ (or equivalently if $\hat{\theta}$ does not fall in the rejection region).

Often one can say int eh last case that "we accept the null hypothesis," but I prefer to not use this language. It is fine, but technically, we are not using the hypothesis test to confirm the null hypothesis. The prupose of the test is to check if our data is *statistically significant*, which means to check if it is sufficiently extreme so as to call in to question the veracity of the null hypothesis.

3 Hypothesis test for mean μ, Z-test

We have normal data $X_i \sim N(\mu, \sigma^2)$ thus we know the sample mean is normally distributed $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

If the data are not normal, then the CLT still allows us to assume $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ approximately, say if $n \geq 30$.

If the data are normal but we do not know the variance σ^2, then we know that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = t \sim T(\mu, \frac{\sigma^2}{n})$ for any n. If the data are not normal, we can still use this as a CLT approximation though.

We want to statistically test a claim about the population mean μ.

3.1 2-sided

Null hypothesis: $H_0 : \mu = \mu_0$
Alternative hypothesis: $H_a : \mu \neq \mu_0$

Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$

p-value: $p = P(Z \leq -|z|) + P(Z \geq |z|) = 2P(Z \leq -|z|)$

In R: $p =$ `2*pnorm(-abs(`z`))`

Note that we take the absolute value here of the test statistic z in order to get the left sided tail. Then we can just double that probability to get the p-value.

Use this test if we know the variance σ^2 and if we know the data is normal. If the data are not normal we can use this as long as the sample size is large ($n \geq 30$ rule of thumb). If we do not know the variance, we can substitute the sample variance s^2 in for σ^2 as long as the sample size is large.

3.2 1-sided

Left side Null hypothesis: $H_0 : \mu \geq \mu_0$ (often still written as $H_0 : \mu = \mu_0$)
Alternative hypothesis: $H_a : \mu < \mu_0$

Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$

p-value: $p = P(Z \leq z)$

In R: $p =$ `pnorm(`z`)`

Note that we are not using an absolute value here. If z is positive, then our p-value will be greater than 50%. In this case, our sample data does not support the alternative hypothesis at all.

Right side Null hypothesis: $H_0 : \mu \leq \mu_0$ (often still written as $H_0 : \mu = \mu_0$)
Alternative hypothesis: $H_a : \mu > \mu_0$

Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$

p-value: $p = P(Z \geq z) = 1 - P(Z < z)$

In R: $p =$ `pnorm(-`z`) = 1-pnorm(`z`)`

Note that we are not using an absolute value here, but instead multiply our test statistic by a minus sign. If z is already negative, then our p-value will be greater than 50%. In this case, our sample data does not support the alternative hypothesis at all.

4 Hypothesis test for mean μ, T-test

Generally, a T-test is better than a Z-test since we are often working with data that are not exactly normal, or with smaller sample sizes.

4.1 2-sided

Null hypothesis: $H_0 : \mu = \mu_0$
Alternative hypothesis: $H_a : \mu \neq \mu_0$

Test statistic: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

p-value: $p = P(T_{n-1} \leq -|t|) + P(T_{n-1} \geq |t|) = 2P(T_{n-1} \leq -|t|)$

In R: $p =$ `2*pt(-abs(`t`),`n`-1)`

If our data is stored in a list `x`, then we can do this t-test in R as:

`t.test(x,mu=`μ_0`,alternative="two.sided")`.

For the 2-sided test, we don't actually need to specify the alternative as it is the default: `t.test(x,mu=`μ_0`)`.

4.2 1-sided

Left side Null hypothesis: $H_0 : \mu \geq \mu_0$ (often still written as $H_0 : \mu = \mu_0$)
Alternative hypothesis: $H_a : \mu < \mu_0$

Test statistic: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

p-value: $p = P(T_{n-1} \leq t)$

In R: $p =$ `pt(`t`)`

If our data is stored in a list `x`, then we can do this t-test in R as:

`t.test(x,mu=`μ_0`,alternative="less")`.

Right side Null hypothesis: $H_0 : \mu \leq \mu_0$ (often still written as $H_0 : \mu = \mu_0$)
Alternative hypothesis: $H_a : \mu > \mu_0$

Test statistic: $t = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$

p-value: $p = P(T_{n-1} \geq t) = 1 - P(T_{n-1} < t)$

In R: $p =$ `pt(-`t`) = 1-pt(`t`)`

If our data is stored in a list `x`, then we can do this t-test in R as:

`t.test(x,mu=`μ_0`,alternative="greater")`.

5 A note about α

The level of significance α needs to be chosen carefully. As already meantioned, $\alpha = 0.05$ is most commonly used, but depending on the application and circumstance, you may wish to set it higher or lower.

Types of errors We do not know with absolute certainty if the null hypothesis is true or not. If it is true and we reject it, we have made a *type I error*. If it is false and we fail to reject it, we have made a *type II error*. So setting α higher will increase our chances of making a type I error. Setting it lower will generally increase our chance of making a type II error.

Consider the example of a construction material that is supposed to withstand a particular stress or force. $H_0 : \mu \geq \mu_0$ and $H_a : \mu < \mu_0$. We may wish to err on the side of caution and reject an acceptable batch of this material rather than use an unacceptable batch in a construction application that might put lives in danger. That being said, there is a cost aspect to consider as well. Rejecting too many materials may increase costs too much, but accepting materials that eventually fail will also increase costs (due to lawsuits and/or insurance claims). There is a careful balance to strike which will require a very careful mathematical analysis that is beyond the scope of this course.

Generally, if you are ok with rejecting a null hypothesis when it is true, but do not want to accept it when it is false, increase α, say to $\alpha = 0.1$ (but $\alpha < 0.5$ is required). If you only want to reject the null hypothesis when it really is false, then you can lower α, say to $\alpha = 0.01$ or even lower.

6 Hypothesis test for proportion p

Null hypothesis: $H_0 : p = p_0$
Alternative hypothesis: $H_a : p \neq p_0$

Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

And perform the test just as above for mean μ.

7 Hypothesis test for variance σ^2

Null hypothesis: $H_0 : \sigma^2 = \sigma_0^2$
Alternative hypothesis: $H_a : \sigma^2 \neq \sigma_0^2$

Test statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

p-value: $p = \min\{P(\chi^2_{n-1} \leq \chi^2), P(\chi^2_{n-1} \geq \chi^2)\} =$ `min(pchisq(`χ^2`,n-1), 1-pchisq(`χ^2`,n-1))`.

If $s^2 < \sigma_0^2$, then $p = 2P(\chi^2_{n-1} \leq \chi^2) =$ `2*pchisq(`χ^2`,n-1)`.

If $s^2 > \sigma_0^2$, then $p = 2P(\chi^2_{n-1} \geq \chi^2) =$ `2*(1-pchisq(`χ^2`,n-1))`.

7.1 1-sided, less

Null hypothesis: $H_0 : \sigma^2 = \sigma_0^2$
Alternative hypothesis: $H_a : \sigma^2 < \sigma_0^2$

Test statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

$p = P(\chi^2_{n-1} \leq \chi^2) =$ `pchisq(`χ^2`,n-1)`.

7.2 1-sided, greater

Null hypothesis: $H_0 : \sigma^2 = \sigma_0^2$
Alternative hypothesis: $H_a : \sigma^2 > \sigma_0^2$

Test statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$

$p = P(\chi^2_{n-1} \geq \chi^2) =$ `1-pchisq(`χ^2`,n-1)`.

8 Hypothesis test for difference in means $\mu_1 - \mu_2$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$
Alternative hypothesis: $H_a : \mu_1 - \mu_2 \neq \Delta_0$

Test statistic: $t = \dfrac{(\overline{x}_1 - \overline{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

p-value: $p = P(T_\nu \leq -|t|) + P(T_\nu \geq |t|) = 2P(T_\nu \leq -|t|)$

In R: $p =$ `2*pt(-abs(`t`),`ν`)`

where the degrees of freedom are

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Round ν down.

If our data is stored in lists `x` and `y`, then we can do this t-test in R as:

`t.test(x,y,mu=`Δ_0`)`.

8.1 1-sided, less

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$
Alternative hypothesis: $H_a : \mu_1 - \mu_2 < \Delta_0$

Test statistic: $t = \dfrac{(\mu_1 - \mu_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

p-value: $p = P(T_\nu \leq -|t|)$

In R: $p =$ `pt(-abs(`t`),`ν`))`

If our data is stored in lists `x` and `y`, then we can do this t-test in R as:

`t.test(x,y,mu=`Δ_0`,alternative="less")`.

8.2 1-sided, greater

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$
Alternative hypothesis: $H_a : \mu_1 - \mu_2 > \Delta_0$

Test statistic: $t = \dfrac{(\mu_1 - \mu_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

p-value: $p = P(T_\nu \leq -|t|)$

In R: $p =$ `pt(-abs(`t`),`ν`))`

If our data is stored in lists `x` and `y`, then we can do this t-test in R as:

`t.test(x,y,mu=`Δ_0`,alternative="greater")`.

9 Hypothesis test for difference in proportions $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_1) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Null hypothesis: $H_0 : p_1 - p_2 = \Delta_0$
Alternative hypothesis: $H_a : p_1 - p_2 \neq \Delta_0$

Test statistic: $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

p-value: $p = P(Z \leq -|z|) + P(Z \geq |z|) = 2P(Z \leq -|z|)$

In R: $p =$ `2*pnorm(-abs(`z`))`

This test is only valid if we have sufficiently large samples and enough of both successes and failures: $n_i\hat{p}_i > 5$ and $n_i(1-\hat{p}_i) > 5$ for $i = 1, 1$.

This test can also be made 1-sided as well.

If $\Delta_0 = 0$, then it is better to use the test statistic:

Test statistic: $z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

where the pooled proportion is

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

10 Chi-squared goodness of fit test

Let's say that we have a collection of classes and a hypothesized proportion for each class. For example Mars Corporation claims to manufacture bags of M&Ms with a specific proportion of each color, and we count the number of M&Ms of each color in a certain bag, then we can test whether or not the claimed proportions fit our specific bag of M&Ms.

Null hypothesis: H_0 : the propotion in class i is p_i for $i = 1, 2, \ldots, k$
Alternative hypothesis: H_a : at least one of the proportions in incorrect

Test statistic: $c = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{k} \frac{(n\hat{p}_i - np_i)^2}{np_i}$

p-value: $p = P(\chi^2_{k-1} \geq c) =$ `1-pchisq(c,df=k-1)`

Example: Suppose it is claimed that a particular population has blood type proportions given by:

blood type	A	B	AB	O
proportion	0.4	0.45	0.1	0.05

Suppose a local medical clinic tests 100 people and finds:

blood type	A	B	AB	O
observed counts	43	40	11	6

The expected counts are $40, 45, 10, 5$ respectively. The test statistic is
$c = \frac{(43-40)^2}{40} + \frac{(40-45)^2}{45} + \frac{(11-10)^2}{10} + \frac{(6-5)^2}{5} = \frac{389}{360} \approx 1.08$

The p-value is $p =$ `1-pchisq(1.08,3)` ≈ 0.78 thus at any significance level below 0.78, we would not reject the null hypothesis. At the typical level of $\alpha = 0.05$, we definitely do not reject H_0. Our data is not that unusual under the assumption that the null hypothesis is true.