

## Hypothesis Tests

### What is a hypothesis?

In statistics, a hypothesis is a statement about a population. Given data from the population, we can assess whether the data support the hypothesis.

For example, the following are statistical hypotheses:

- The population mean is a positive number.
- The median of population A is greater than the median of population B.
- The probability of observing a value greater than 10 is less than  $1/2$ .
- The variance of population A is less than the variance of population B.
- The population is right-skewed.

With a finite data set, it is never possible to be certain about the truth of a hypothesis. Hypothesis testing is a means to provide a quantitative summarization of the evidence in a given set of data in favor of or against a hypothesis.

### Hypothesis testing as a decision problem

If a hypothesis is deemed inconsistent with the data it is “rejected,” otherwise it is “accepted.” In order to make this decision, we need a test statistic to quantify how well the data fit the hypothesis.

If the data are  $X_1, \dots, X_n$ , the test statistic is a function  $T(X_1, \dots, X_n)$  that compresses all the relevant information in the data about the hypothesis into a single number. The test statistic is usually constructed so that values of  $T$  close to zero indicate strong consistency between the data and the hypothesis in question, whereas values of  $T$  far from zero indicate poor consistency between the data and hypothesis.

To make a decision based on  $T$ , a “critical value”  $T_0$  is specified, so that the hypothesis is rejected under the following circumstances:

- *Two-sided alternative*: reject if  $|T| > T_0$
- *Right-tailed alternative*: reject if  $T > T_0$
- *Left-tailed alternative*: reject if  $T < -T_0$

These three situations will be clarified below.

### Asymmetry in hypothesis testing

In most research investigations, one begins by assuming that the relationship under study does not exist. This assumption is the “null hypothesis.” If in fact the relationship is real, the “alternative hypothesis” is true.

For example, if a study is being carried out to assess whether a newly developed drug is effective, the null hypothesis would be that it is not effective. The alternative hypothesis would be that it is effective.

Viewing a hypothesis test as a decision problem, there are two possible correct outcomes and two possible incorrect outcomes, as shown in the following table.

		Truth	
		Null	Alternative
Decision	Null	True negative	False negative
	Alternative	False positive	True positive

A “negative” is a decision in favor of the null hypothesis and a “positive” is a decision in favor of the alternative hypothesis. Either a negative decision or a positive decision can be true or false, as indicated in the table.

Returning to the drug example, the false decisions are as follows.

- A false negative occurs if the drug is truly effective but is falsely deemed ineffective. The cost of this mistake is that patients do not benefit from the therapeutic effect of the drug.
- A false positive occurs if the drug is ineffective, but is falsely deemed to be effective. The cost of this mistake is that patients may be given an ineffective drug, when other alternatives may be available.

Hypothesis testing problems are usually set up so that a false positive is a more costly mistake than a false negative. Therefore, the probability of a false positive occurring is bounded by a constant  $\alpha$  called the “level” of the test. The level of a test determines the critical value. For example, for a two-sided test,

$$P_{\text{null}}(|T| > T_0) = \alpha.$$

If we know the sampling distribution of  $T$ , we can solve this equation for  $T_0$ .

## Hypothesis testing without decisions

An different approach to hypothesis testing is to quantify the evidence against the null hypothesis without making an explicit decision. Suppose  $T_{\text{obs}}$  is the test statistic calculated for the observed data, and let  $T$  represent the sampling distribution of the test statistic under the null hypothesis. The “p-value” is the probability of getting as much or more evidence against the null as is represented by  $T_{\text{obs}}$ . Specifically, the p-value is

- $P(|T| > T_{\text{obs}})$  (two-sided test)
- $P(T > T_{\text{obs}})$  (one-sided right-tailed test)
- $P(T < -T_{\text{obs}})$  (one-sided left-tailed test)

The p-value is the probability of observing as much or more evidence against the null hypothesis as was actually observed, when the null hypothesis is assumed to be true. It is the likelihood that the apparent evidence against the null hypothesis is due to chance.

The p-value can be reported without deciding whether to reject the null hypothesis. Advantages of doing this are that it is less formal, it provides more quantitative information about the strength of evidence against the null (compared to a yes/no decision), and if there is a subsequent need to make a decision at a given level (e.g.  $\alpha = 0.05$ ), the p-value is sufficient for doing so.

## Examples:

### One-sample location tests

In a one-sample location test, independent and identically distributed data  $X_1 \dots, X_n$  are observed, and the null hypothesis is  $EX_i = 0$ .

The test statistic is

$$\sqrt{n}\bar{X}/\hat{\sigma},$$

when  $\sigma^2 = \text{var } X_i$  is not known, or

$$\sqrt{n}\bar{X}/\sigma$$

when  $\sigma^2$  is known.

When  $\sigma^2$  is known, the test statistic has approximately a standard normal distribution under the null hypothesis. Thus the p-value is

$$1 - P(Z < \sqrt{n}\bar{X}/\sigma).$$

where  $Z$  is standard normal. When  $\sigma^2$  is not known and the data are normal,  $\sqrt{n}\bar{X}/\hat{\sigma}$  follows a  $t$  distribution with  $n - 1$  degrees of freedom under the null hypothesis. Thus the p-value is

$$1 - P(T < \sqrt{n}\bar{X}/\hat{\sigma}),$$

where  $T$  follows the  $t_{n-1}$  distribution.

The following simulation study checks the of p-values for the one-sample location test when the variance is known. The goal is to directly confirm that null draws from the sampling distribution of the test statistic exceed the actual test statistic proportion  $p$  of the time, where  $p$  is the p-value.

```

## Column 1 is the p-value, column 2 is the proportion of null test
## statistics exceeding the actual test statistic.
Q <- array(0, c(1000,2))

for (r in (1:1000))
{
  ## Generate an alternative data set.
  X <- runif(1) + rnorm(20)

  ## Get the p-value, assuming that the variance is known to be 1.
  M1 <- mean(X)
  p <- 1-pnorm(sqrt(20)*M1)

  ## Check the accuracy of the p-value using simulated null data.
  Z <- array(rnorm(20*1000), c(20,1000))
  M2 <- colMeans(Z)

  Q[r,1] <- p
  Q[r,2] <- mean(M2 > M1)
}

```

We should also study how the test performs under null data. It is a fact that under null data, the p-values follow a uniform distribution. This means that 1% of the p-values are less than 0.01, 5% of the p-values are less than 0.05, and so on. The following simulation assesses whether the one-sample location test with known variance behaves consistently with this fact.

```

## Storage for 10000 null p-values.
Q <- array(0, 10000)

for (r in (1:10000))
{
  ## Generate a null data set.
  X <- rnorm(20)

  ## Get the p-value, assuming that the variance is known to be 1.
  M1 <- mean(X)
  Q[r] <- 1-pnorm(sqrt(20)*M1)
}

```

To evaluate the results, make a histogram of the values in Q.

## Two-sample location tests

Suppose we want to compare the expected values for two populations. For example, population A may represent the treatment responses of people treated with a newly developed drug, while population B represents the treatment responses of people treated with the conventional drug.

Suppose we observe a sample  $X_1, \dots, X_n$  from population A and a sample  $Y_1, \dots, Y_m$  from population B. To compare treatment responses in the two groups, the null hypothesis will be

$$EX = EY$$

and the alternative hypothesis will be  $EX > EY$  (assuming that greater values correspond to better response).

If the variances of populations A and B, denoted  $\sigma_A^2$  and  $\sigma_B^2$ , are known, the “Z-statistic”

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_A^2/n + \sigma_B^2/m}}$$

has mean zero and variance one under the null hypothesis. If the sample sizes are large, or if the data are approximately normal,  $T$  approximately has a standard normal distribution under the null hypothesis.

If the variances are unknown, the plug-in version of the Z-statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}_A^2/n + \hat{\sigma}_B^2/m}}$$

may be used. If the sample size is not too small, the plug-in version of  $T$  is approximately standard normal, but if the sample size is small it may be quite far from being standard normal.

If the sample size is small, and the population variances are unknown, and the data are thought to be approximately normal, the two-sample t-statistic:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(m+n)/mn}}$$

can be used, where

$$S_p^2 = \left( \sum_i (X_i - \bar{X})^2 + \sum_i (Y_i - \bar{Y})^2 \right) / (m + n - 2)$$

is the “pooled variance estimate.” Under the null hypothesis,  $T$  has a  $t_{n+m-2}$  distribution.

The following simulation study looks at the accuracy of p-values produced by the plug-in Z-statistic and the t-statistic.

```
## Column 1 is the plug-in Z-statistic p-value, column 2 is the
## t-statistic p-value, column 3 is the proportion of 1000 null
## plug-in Z-statistics exceeding the actual plug-in Z-statistic, and
## column 4 is the proportion of 1000 null t-statistics exceeding the
## actual t-statistic.
Q <- array(0, c(1000,4))

## Sample sizes for the two groups.
m <- 10
```

```

n <- 5

## Do 1000 replications.
for (r in (1:1000))
{
  ## Population means and variances for populations A and B.
  ## The minimum separation between the means is 1.
  M <- c(2, runif(1))
  V <- 2*runif(2)

  ## Generate the sample data.
  X <- M[1] + sqrt(V[1])*rnorm(n)
  Y <- M[2] + sqrt(V[2])*rnorm(m)

  ## Get the plug-in Z-statistic p-value.
  MD <- mean(X) - mean(Y)
  VX <- var(X)
  VY <- var(Y)
  TZ <- MD / sqrt(VX/n + VY/m)
  Q[r,1] <- 1-pnorm(TZ)

  ## Get the T-statistic p-value.
  Sp2 <- ((n-1)*VX + (m-1)*VY) / (n+m-2)
  TT <- MD / sqrt(Sp2*(m+n)/(m*n))
  Q[r,2] <- 1-pt(TT, n+m-2)

  ## Simulate 1000 null data sets.
  X <- array(rnorm(n*1000), c(n,1000))
  Y <- array(rnorm(m*1000), c(m,1000))

  ## Calculate plug-in Z-statistics for the simulated data sets.
  MD <- colMeans(X) - colMeans(Y)
  VX <- apply(X, 2, var)
  VY <- apply(Y, 2, var)

```

```

TZS <- MD / sqrt(VX/n + VY/m)

## Calculate t-statistics for the simulated data sets.
Sp2 <- ((n-1)*VX + (m-1)*VY) / (n+m-2)
TTS <- MD / sqrt((m+n)*Sp2/(m*n))

## Get the proportion of null test statistics exceeding the actual
## test statistic, for both the plug-in Z statistic (column 3) and
## the t-statistic (column 4).
Q[r,3] <- mean(TZS > TZ)
Q[r,4] <- mean(TTS > TT)
}

```

This program can be used to address a number of important questions:

- How much do the t-statistic p-values differ from the plug-in Z-statistic p-values? To start with this, try

```

plot(Q[,1:2])
lines(cbind(c(0,1), c(0,1)))

```

- How much more accurate overall are the t-statistic p-values compared to the plug-in Z-statistic p-values? To start with this, try

```

plot(Q[,c(1,3)])
lines(cbind(c(0,1), c(0,1)))
x11()
plot(Q[,c(2,4)])
lines(cbind(c(0,1), c(0,1)))

```

- Since we are more concerned about small p-values being accurate, it is better to compare the p-values in a log/log plot:

```

plot(log(Q[,c(1,3)]))
lines(cbind(c(-10,0), c(-10,0)))

```

```
x11()  
plot(log(Q[,c(2,4)]))  
lines(cbind(c(-10,0), c(-10,0)))
```

- How is the discrepancy between the two approaches affected by the total sample size  $m + n$ , and by the discrepancy  $m - n$  between the sample sizes?
- How is the discrepancy between the two approaches affected by the presence of unequal variances in the two populations?
- How do the two methods compare if the data are not normal?

## Paired tests

Suppose we observe paired data

$$(X_i, Y_i)$$

$i = 1, \dots, n$ . For example, the values in the pair might represent before-treatment and after-treatment measurements for a patient in a medical trial. Our interest is in testing whether the treatment effect is nonzero.

We might just ignore the pairing of the data, and proceed as in a two-sample location test. However this ignores individual-specific factors. For example, a person who is more sick to begin with will have poorer measures both before and after treatment. But such a person could have an equal, or greater treatment response as a relatively healthier person.

For paired data, it is almost always preferable to work with differences

$$D_i = Y_i - X_i.$$

The hypothesis of no treatment effect can now be expressed  $ED_i = 0$ . This is a one sample location problem, as discussed above.

## Permutation tests

If the data are not normal and the sample size is small, neither the Z-statistic nor t-statistic is easy to calibrate – that is, p-values from the standard normal or  $t_{n+m-2}$  distributions may not be very accurate.

A completely different approach to calculating p-values is to work only with the available data, without using any model for the data. In order to produce a p-value, we need to generate many replicated data sets from an appropriate null distribution. If we are able to do this, the proportion of test statistic values for the simulated null data sets that exceed the actual test statistic value can be used as a p-value.

For the two-sample location problem, we observe data sets  $X_i$  and  $Y_i$ , and use one of the test statistics discussed above for assessing whether  $EX_i = EY_i$ . We'll use the plug-in Z-statistic here:

$$T_{\text{obs}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}_X^2/m + \hat{\sigma}_Y^2/n}}.$$

We need to construct a null-distribution for calculating a p-value,  $P(|T| > T_{\text{obs}})$ . To do this, we randomly reassign the observed  $X$  and  $Y$  values to two groups having the same sizes as the actual groups.

For example, if the actual  $X$  data are 1, 3, 4 and the actual  $Y$  data are 2, 2, 1, 2, we first pool everything together, yielding

$$1, 1, 2, 2, 2, 3, 4.$$

Next we randomly permute the values, yielding (for example)

$$3, 1, 2, 4, 2, 2, 1$$

then split these values into artificial  $X$  and  $Y$  sets of the same size as the actual  $X$  and  $Y$  sets. The artificial  $X$  set is 3, 1, 2 and the artificial  $Y$  set is 4, 2, 2, 1.

For each such constructed data set, a null test statistics  $T_i$  is calculated. The two-sided permutation p-value is then simply

$$\sum_i I(|T_i| > |T_{\text{obs}}|)/N,$$

where  $N$  is the number of permuted sets that were constructed, and  $I(\cdot)$  is the indicator function yielding one if its argument is true, and zero otherwise.

If we have data in vectors  $X$  and  $Y$ , then the following code gives the permutation test p-value for the null hypothesis that the population means for the two populations are equal.

```
## The sample sizes.
n <- length(X)
m <- length(Y)

## Calculate the Z-statistic for the actual data.
mx <- mean(X)
my <- mean(Y)
vx <- var(X)
vy <- var(Y)
T <- (mx - my) / sqrt(vx/n + vy/n)

## Merge all the data together.
Z <- c(X, Y)

## Get 1000 Z-statistics for permuted data.
TR <- array(0, 1000)
for (r in (1:1000))
{
  ## Generate a random permutation.
  ii <- order(runif(m+n))

  ## Construct x and y data sets by random reassignment.
  x <- Z[ii[1:n]]
  y <- Z[ii[(n+1):(n+m)]]
}
```

```

## Calculate the Z-statistic for the reassigned data.
mx <- mean(x)
my <- mean(y)
vx <- var(x)
vy <- var(y)
TR[r] <- (mx - my) / sqrt(vx/n + vy/n)
}

## Two-sided p-value.
pv2 <- mean(abs(TR) > abs(T))

```

To get a one-sided right-tailed p-value, replace the final line of code with

```
pvr <- mean(TR > T),
```

and to get a one-sided left-tailed p-value, replace the final line of code with

```
pvl <- mean(TR < T).
```

### *Performance of permutation tests*

Just as with the bootstrap, the permutation test is an appealing idea, but we should check that it actually works. The following simulation checks the distribution of permutation test p-values under one possible null distribution, in which the two populations being compared are normal with mean zero, but may have different variances (the variances are generated independently from a standard exponential distribution). In interpreting the results of this simulation, remember that p-values from a null distribution must follow a uniform distribution.

```

## The sample sizes.
n <- 10
m <- 10

## Generate permutation-test p-values for 1000 null data sets.
pv <- array(0, 1000)
for (j in (1:1000))

```

```

{
## Variances for the two populations.
V <- rexp(2)

## Generate the 'actual' data.
X <- sqrt(V[1])*rnorm(n)
Y <- sqrt(V[2])*rnorm(m)

## Calculate the Z-statistic for the actual data.
mx <- mean(X)
my <- mean(Y)
vx <- var(X)
vy <- var(Y)
T <- (mx - my) / sqrt(vx/n + vy/n)

## Merge all the data together.
Z <- c(X, Y)

## Get 1000 Z-statistics for permuted data.
TR <- array(0, 1000)
for (r in (1:1000))
{
## Generate a random permutation.
ii <- order(runif(m+n))

## Construct x and y data sets by random reassignment.
x <- Z[ii[1:n]]
y <- Z[ii[(n+1):(n+m)]]

## Calculate the Z-statistic for the reassigned data.
mx <- mean(x)
my <- mean(y)
vx <- var(x)
vy <- var(y)
}
}

```

```

    TR[r] <- (mx - my) / sqrt(vx/n + vy/n)
  }

  ## Two-sided p-value.
  pv[j] <- mean(abs(TR) > abs(T))
}

```

The use of unequal variances in the previous simulation is important. One potential issue with the permutation test for the equality of means is that it implicitly assumes that all other aspects of the two distributions are equal (for example, the variance). If the true variances are unequal, but the simulated null data sets have equal variances, this could potentially lead to systematic inaccuracies in the p-values.

## Power

The power of a hypothesis test is the probability of making a decision in favor of the alternative hypothesis when the alternative hypothesis is true. The power depends on the sample size, the type of statistic being used, the noise level of the data, and the structure of the alternative hypothesis. While the sample size and statistic being used are known in advance, the structure of the alternative distribution is not. For example, in the two sample location problem, the alternative hypothesis is that the population means in the two populations being compared differ. The power depends on the amount by which they differ, as well as the variances in the two populations, all of which are unknown.

A power analysis is a systematic exploration of the power of a hypothesis test under various plausible scenarios for the alternative distribution. In certain cases, the power can be derived mathematically. For example, in the two-sample location problem, suppose that the data  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$  are normal with means  $\mu_X = \mu_Y$ , and variances  $\sigma_X^2$  and  $\sigma_Y^2$ . Suppose further that the variances are known. A one sided test using the Z-statistic rejects the null hypothesis if

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} > 1.64.$$

Now suppose that the alternative hypothesis is that  $\mu_X - \mu_Y = \delta > 0$ . To begin deriving the power, note that

$$P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} > 1.64\right) = P\left(\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} > 1.64 - \delta/\sqrt{\sigma_X^2/n + \sigma_Y^2/m}\right),$$

where

$$\frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$$

is standard normal under the alternative hypothesis. Thus the power is

$$1 - P(Z < 1.64 - \delta/\sqrt{\sigma_X^2/n + \sigma_Y^2/m}),$$

where  $Z$  is standard normal. Note how the power changes as various quantities are changed:

- Larger values of  $\delta$  lead to more power.
- Larger variances (i.e.  $\sigma_X^2$  and  $\sigma_Y^2$ ) lead to less power.
- Larger sample sizes lead to more power.
- More stringent control of the false positive probability requires replacing 1.64 with a larger number, leading to less power.

In more complicated settings deriving the power mathematically is difficult or impossible. However a simulation can always be performed to estimate the power. For example, suppose we are doing a one-sided two-sample location test using the t-statistic with unknown population variances for the  $X$  and  $Y$  populations. The following simulation estimates the power for several possible values of  $\delta$ , the common variance of  $X$  and  $Y$ , and the sample size.

```
## Consider these values for EX-EY.
for (d in c(0.5,1,1.5))
{
```

```

## Consider these sample sizes (for both the X and Y sample).
for (n in c(10,20,30))
{
  ## Consider these values for var(X) and var(Y) (we are assuming
  ## that the variances are equal here).
  for (v in c(1,2,3))
  {
    ## Generate 1000 data sets from the appropriate alternative
    ## distribution.
    X <- array(d + sqrt(v)*rnorm(n*1000), c(1000,n))
    Y <- array(sqrt(v)*rnorm(n*1000), c(1000,n))

    ## Calculate T-statistics for the simulated data sets.
    MX <- rowMeans(X)
    MY <- rowMeans(Y)
    VX <- apply(X, 1, var)
    VY <- apply(Y, 1, var)
    Sp2 <- ((n-1)*VX + (n-1)*VY) / (2*n-2)
    TS <- (MX-MY) / sqrt(2*Sp2/n)

    ## This is an estimate of the power.
    pw <- mean(TS > qt(0.95, 2*n-2))

    ## Print the result.
    print(c(d, n, v, pw))
  }
}
}

```