07 Hypothesis Testing

What is a Hypothesis?

- A hypothesis is a statement about a population. Given data from the population, we can assess whether the data support the hypothesis.
- Elements of a statistical test include a null and alternative hypothesis, assumption checking, a test statistic, a p-value, a decision, and a conclusion.
- The choice of the test statistic depends on the distribution of the population from which the data come from and the hypotheses being considered.

Null and Alternative Hypotheses

- The null hypothesis, H_0, represents the status quo or statement of no effect.
  - It is generally the model that the experimenter would like to replace.
- The alternative hypothesis, $H_a (H_1)$, usually represents the experimenter's new model, what the experimenter would like to support.
  - It may be a denial of the null hypothesis (two-sided test).
  - It may specify a direction of interest (one-sided test).
- With a finite data set, it is never possible to be certain about the truth of the null hypothesis.

Test Statistics

- Hypothesis testing provides a quantitative summary of the evidence in a given data in favor of or against a hypothesis.
- If the null hypothesis is inconsistent with the data, it is "rejected," otherwise it is "not rejected."
- In order to decide whether to reject or accept the null, a test statistic is need to assess how well the data fit the hypothesis.
- If the data are $X_1, \ldots, X_n$, the test statistic is a function $T(X_1, \ldots, X_n)$ that compresses all the relevant information in the data about the hypothesis into a single number.
- A test statistic should be constructed so that values of T close to zero indicate that the data strongly agree with the null hypothesis, and values of T far from zero indicate poor agreement between the data and the null hypothesis.

Hypothesis Testing as a Decision Problem

- 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 $T < t_0$
- 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.

Decision Results

- 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.
  - Reject H_0 $\approx$ positive $\approx$ significant
  - Fail to reject H_0 $\approx$ negative $\approx$ not significant

Drug Example

- For example, a study is being carried out to assess whether a newly developed drug is effective or not.
  - The null hypothesis would be that it is not effective.
  - The alternative hypothesis would be that it is effective.
- In 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 are given an ineffective drug, when effective alternatives may be available.

Significance Level

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

  $$P(|T| > t_0) = \alpha.$$

  when H_0 is true.
- If we know the sampling distribution of T, we can solve this equation for t_0.

Computing the p-value

- A different approach to hypothesis testing is to quantify the evidence against the null hypothesis without making an explicit decision.
- Suppose T_{obs} is the test statistic calculated from 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_{obs}.
- The p-value can be calculated as follows:
  - $P(|T| > T_{\text{obs}})$ for two-sided test
  - $P(T > T_{\text{obs}})$ for one-sided right-tailed test
  - $P(T < T_{\text{obs}})$ for one-sided left-tailed test

Statistical Decision based on p-value

- Common values of significance level α are
  - 0.01, 0.05, and 0.10.
- The decision is made to reject H_0 if the p-value is less than or equal to α. If we reject the null hypothesis, the results of the test are said to be statistically significant at the level α.
- A "significant" result in the statistical sense does not necessarily imply an "important" result. It means simply that such a difference from the null hypothesis is "not very likely to happen just by chance."

Test Errors

- There are two types of errors in hypothesis testing.
- If the null hypothesis is true but the decision is to reject H_0, then a Type I error is said to have occurred.

$$\text{Type I error rate} = P(\text{reject } H_0 | H_0 \text{ is true})$$

- Failing to reject H_0 when the alternative hypothesis is true is called a Type II error.

$$\text{Type II error rate} = P(\text{Not reject } H_0 | H_a \text{ is true})$$

- If the null hypothesis is true, the significance level α is also the probability of a Type I error.
- The probability of a Type II error is denoted by β.

Test Error Rates and Decision Results

- There are connection between test errors and decision outcomes
  - False positive rate = Type I error rate (α)
  - False negative rate = Type II error rate (β)
  - True positive rate (sensitivity) = statistical power $(1-\beta)$
  - True negative rate (specificity) $= 1-\alpha$

		Truth	
		Null	Alternative
Decision	Null	True negative (TN)	False negative (FN)
	Alternative	False positive (FP)	True positive (TP)

  - TPR (sensitivity) = TP/(FN + TP)
  - TNR (specificity) = TN/(TN + FP)
  - FPR (α) $= 1-$ TNR = FP/(TN + FP)
  - FNR (β) $= 1 =$ TPR = FN/(FN+TP)

Statistical Power

- The power of a test measures its ability to detect an alternative hypothesis when it is true.
- Power against a particular alternative is calculated as the probability that the test will reject H_0 when the alternative hypothesis is true and thus represented by $1 - \beta$.
- Decisions and errors

H_0 true	$\Rightarrow$	Reject H_0	$\Rightarrow$	Type I error
H_0 true	$\Rightarrow$	Fail to reject H_0	$\Rightarrow$	Correct decision
H_0 false	$\Rightarrow$	Reject H_0	$\Rightarrow$	Correct decision
H_0 false	$\Rightarrow$	Fail to reject H_0	$\Rightarrow$	Type II error

Statistical 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.
- For a two-sided test, this is

$$P(|T| > t_0 | H_0 \text{ is false}).$$

- The power depends on:
  - The sample size.
  - The statistic being used.
  - The significance level.
  - The alternative distribution.

One-sample Test for Population Mean μ

- The null and alternative hypotheses are
  - $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$
  - $H_0 : \mu \leq \mu_0$ vs. $H_a : \mu > \mu_0$ (right)
  - $H_0 : \mu \geq \mu_0$ vs. $H_a : \mu < \mu_0$ (left)
- If σ is known, the test statistic is

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

- If σ is unknown, the test statistic is

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim T_{n-1}$$

where $\hat{\sigma}$ is a sample standard deviation.

Statistical Decision for One-sample Test

- Statistical decision for two-side test at a level α
  - Z-test: reject H_0 if $|Z| > z_{1-\alpha/2}$
  - T-test: reject H_0 if $|T| > t_{n-1,1-\alpha/2}$
  - Reject H_0 if p-value $< \alpha$
- Statistical decision for one-side test at a level α
  - Z-test (right): reject H_0 if $Z > z_{1-\alpha}$
  - Z-test (left): reject H_0 if $Z < z_{\alpha}$
  - T-test (right): reject H_0 if $T > t_{n-1,1-\alpha}$
  - T-test (left): reject H_0 if $T < t_{n-1,\alpha}$
  - Reject H_0 if p-value $< \alpha$

```
## Two side critical value of Z test
qnorm(1 - alpha/2)

## Two side critical value of T test
qt(1 - alpha/2, df)
```

One-sample Test Example

- The following simulation studies conduct one-sample Z-test for

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_a : \mu \neq 0$$

- When `mu = null`, the total number of rejections among `nrep` simulation replications indicates type I error rate computation, since the null is actually true.
- When `mu != null`, the total number of rejections among `nrep` simulation replications indicates statistical power computation, since the null is actually false.
- Note that the total number of p-values less than `alpha` among `nrep` simulation replications is equivalent to the total number of rejections.

```
set.seed(1234)
mu <- 0                      ## population mean
sig <- 1                     ## population standard deviation
alpha <- c(0.01, 0.05, 0.1)  ## significance level
n <- 20                      ## sample size
nrep <- 1e4                  ## simulation replications
null <- 0                    ## null value

out <- matrix(0, length(alpha), 2)
colnames(out) <- c("p.value", "test.stat")
rownames(out) <- alpha
for (i in 1:length(alpha)) {
     X <- matrix(rnorm(nrep*n, mean=mu, sd=sig), nrep, n)
     T <- apply(X, 1, function(t) (mean(t)-null)/(sig/sqrt(n)))
     pval <- (1 - pnorm(abs(T)))*2
     test <-  abs(T) > qnorm(1 - alpha[i]/2)
     out[i, 1] <- mean(pval < alpha[i])
     out[i, 2] <- mean(test)
}
out
```

```
set.seed(1234)
mu <- 0.5                    ## mu=0.5 != 0
sig <- 1
alpha <- c(0.01, 0.05, 0.1)
n <- 20                      ## sample size = 20
nrep <- 1e4
null <- 0

out <- matrix(0, length(alpha), 2)
colnames(out) <- c("p.value", "test.stat")
rownames(out) <- alpha
for (i in 1:length(alpha)) {
     X <- matrix(rnorm(nrep*n, mean=mu, sd=sig), nrep, n)
     T <- apply(X, 1, function(t) (mean(t)-null)/(sig/sqrt(n)))
     pval <- (1 - pnorm(abs(T)))*2
     test <-  abs(T) > qnorm(1 - alpha[i]/2)
     out[i, 1] <- mean(pval < alpha[i])
     out[i, 2] <- mean(test)
}
out
```

```
set.seed(1234)
mu <- 1                      ## mu = 1
sig <- 1
alpha <- c(0.01, 0.05, 0.1)
n <- 20                      ## sample size = 20
nrep <- 1e4
null <- 0

out <- matrix(0, length(alpha), 2)
colnames(out) <- c("p.value", "test.stat")
rownames(out) <- alpha
for (i in 1:length(alpha)) {
     X <- matrix(rnorm(nrep*n, mean=mu, sd=sig), nrep, n)
     T <- apply(X, 1, function(t) (mean(t)-null)/(sig/sqrt(n)))
     pval <- (1 - pnorm(abs(T)))*2
     test <-  abs(T) > qnorm(1 - alpha[i]/2)
     out[i, 1] <- mean(pval < alpha[i])
     out[i, 2] <- mean(test)
}
out
```

```
set.seed(1234)
mu <- 0.5                    ## mu = 0.5
sig <- 1
alpha <- c(0.01, 0.05, 0.1)
n <- 50                      ## sample size = 50
nrep <- 1e4
null <- 0

out <- matrix(0, length(alpha), 2)
colnames(out) <- c("p.value", "test.stat")
rownames(out) <- alpha
for (i in 1:length(alpha)) {
     X <- matrix(rnorm(nrep*n, mean=mu, sd=sig), nrep, n)
     T <- apply(X, 1, function(t) (mean(t)-null)/(sig/sqrt(n)))
     pval <- (1 - pnorm(abs(T)))*2
     test <-  abs(T) > qnorm(1 - alpha[i]/2)
     out[i, 1] <- mean(pval < alpha[i])
     out[i, 2] <- mean(test)
}
out
```

Statistical Power Example

- The following simulation study conducts one-sample T-test for

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_a : \mu \neq 0$$

- The simulation computes statistical power when
  - $X \sim N(\mu, 1)$, where $\mu =$ 0.3, 0.5 and 1
  - The sample size $n =$ 10, 20, 30 and 50
  - The significance level $\alpha =$ 0.01, 0.05 and 0.1
- In this simulation, we denote
  - μ : `mu`
  - n : `nsamp`
  - α : `alp`

```r
set.seed(13579)
mu <- c(0.3, 0.5, 1); nsamp <- c(10, 20, 30, 50)
alp <- c(0.01, 0.05, 0.1); nrep <- 1e4; RE <- NULL
for (i in 1:length(mu)) {
  d <- mu[i]
  for (j in 1:length(nsamp)) {
     n <- nsamp[j]
     for (k in 1:length(alp)) {
       a <- alp[k]
       X <- matrix(rnorm(n*nrep, mean=d, sd=1), nrep, n)
       MX <- apply(X, 1, mean)
       SX <- apply(X, 1, sd)
       T <- MX/(SX/sqrt(n))
       pw <- mean(abs(T) > qt(1 - a/2, df = n - 1))
       RE <- rbind(RE, c(d, n, a, pw))
     }
  }
}
```

```
colnames(RE) <- c("mu", "n", "alpha", "power")
RE

par(mfrow=c(1,3))
for (i in 1:length(mu)) {
    Z <- RE[RE[,1]==mu[i],-1]
    interaction.plot(Z[,1], Z[,2], Z[,3], type="b", pch=19,
                     col=c(1,2,4),  ylim=c(0,1),
                     xlab="Sample size (n)", legend=FALSE,
                     ylab="Statistical power",
                     main=paste("mu = ", mu[i]),)
    if (i==1) legend("topleft", c("alpha = 0.01",
                     "alpha = 0.05", "alpha = 0.1"),
                     lty=c(3,2,1), pch=19, col=c(1,2,4),
                     cex=1.2)
}
```

Statistical Power Example

- The following simulation studies compare the powers of Z-test and one-sample T-test,

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad \text{vs.} \quad T = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

  when testing

$$H_0 : \mu = \mu_0 \qquad \text{vs.} \qquad H_a : \mu \neq \mu_0$$

- In this simulation, we fix
  - $\mu_0 = 0$
  - $n = 5, 10, 20, 30, 50, 100$ and 1000
  - $\alpha = 0.05$
- The population distribution is generated from
  - $x_i \sim N(0.5, 1)$, so $\mu = 0.5$
  - $x_i \sim Unif(-1, 1.5)$, so $\mu = 0.25$

```
set.seed(54321)
mu <- 0.5; alp <- 0.05; nrep <- 1e4
nsamp <- c(5, 10, 20, 30, 50, 100, 1000)
POW <- matrix(0, length(nsamp), 2)
rownames(POW) <- nsamp
colnames(POW) <- c("Z-test", "T-test")
for (i in 1:length(nsamp)) {
    n <- nsamp[i]
    X <- matrix(rnorm(n*nrep, mean=mu, sd=1), nrep, n)
    MX <- apply(X, 1, mean)
    SX <- apply(X, 1, sd)
    Z <- MX/(1/sqrt(n))
    T <- MX/(SX/sqrt(n))
    POW[i, 1] <- mean(abs(Z) > qnorm(1-alp/2))
    POW[i, 2] <- mean(abs(T) > qt(1-alp/2, df=n-1))
}
POW
matplot(POW, type="l", col=c(2,4), xaxt="n", xlab="n")
axis(1, at=seq(nsamp), labels=nsamp)
legend("topleft", c("Z", "T"), col=c(2,4), lty=c(1,2), cex=1.2)
```

```r
set.seed(1111)
mu <- c(-1, 1.5); alp <- 0.05; nrep <- 1e4
nsamp <- c(5, 10, 20, 30, 50, 100, 1000)
POW <- matrix(0, length(nsamp), 2)
rownames(POW) <- nsamp
colnames(POW) <- c("Z-test", "T-test")
for (i in 1:length(nsamp)) {
    n <- nsamp[i]
    X <- matrix(runif(n*nrep, mu[1], mu[2]), nrep, n)
    MX <- apply(X, 1, mean)
    SX <- apply(X, 1, sd)
    pop.sd <- sqrt((mu[2]-mu[1])^2/12)
    Z <- MX/(pop.sd/sqrt(n))
    T <- MX/(SX/sqrt(n))
    POW[i, 1] <- mean(abs(Z) > qnorm(1-alp/2))
    POW[i, 2] <- mean(abs(T) > qt(1-alp/2, df=n-1))
}
POW
matplot(POW, type="l", col=c(2,4), xaxt="n", xlab="n")
axis(1, at=seq(nsamp), labels=nsamp)
legend("topleft", c("Z", "T"), col=c(2,4), lty=c(1,2), cex=1.2)
```

Two-sample Tests for Population Mean Difference

- Suppose we want to compare the means of 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 $X_1, \ldots, X_n$ from population A and a sample $Y_1, \ldots, Y_m$ from population B.
- To compare treatment responses in the two groups, the hypothesis will be

$$H_0 : \mu_X = \mu_Y \quad \text{vs.} \quad H_1 : \mu_X > \mu_Y$$

assuming that greater values correspond to better response.

Two-sample Z-test

- If the variances of populations A and B, denoted σ_X^2 and σ_Y^2 are known, the "Z-statistic" can be used.

  $$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_X^2}{m}}}$$

  has a mean of zero and a variance of one under the null hypothesis.

- If the sample sizes are large, or if the data are approximately normal, Z approximately has a standard normal distribution under the null hypothesis.

  $$Z \sim N(0, 1)$$

General Two-sample T-test

- If the variances are unknown, the plug-in version of the T-statistic can be used.

$$T_1 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}}$$

- If the sample size is not too small, the plug-in version of T_1 is approximately standard normal, but if the sample size is small it may be quite far from being standard normal.
- When the population variances $\sigma_X^2 \neq \sigma_Y^2$, it is often assumed that T_1 has a t-distribution with a degree of freedom of $\min(n-1, m-1)$.

Pooled Two-sample T-test

- If the sample size is small, and the population variances are assumed to be equal, and the data are thought to be approximately normal, the pooled two-sample t-statistic can be used.

  $$T_2 = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

  where

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

  is the "pooled variance estimate."
- Under the null hypothesis, T_2 has a t-distribution with a degree of freedom of $n + m - 2$.

Example

- The next simulation study compare the type I error rates of three test statistics.
  - $Z \sim N(0,1)$
  - $T_1 \sim t_{\min(n-1,m-1)}$
  - $T_2 \sim t_{n+m-2}$

```r
set.seed(1111)
nrep <- 1e4
Q <- matrix(0, nrep, 3)

n <- 5
m <- 10
mu_x <-  0
mu_y <-  0
sig_x <-  1
sig_y <-  1
```

```
for (r in 1:nrep) {
    X <- rnorm(n, mean=mu_x, sd=sig_x)
    Y <- rnorm(m, mean=mu_y, sd=sig_y)

    MD <- mean(X) - mean(Y)
    VX <- var(X)
    VY <- var(Y)
    T1 <- MD / sqrt(VX/n + VY/m)

    Sp2 <- ((n-1)*VX + (m-1)*VY) / (n+m-2)
    T2 <- MD / sqrt(Sp2*(1/n+1/m))

    Q[r, 1] <- 1 - pnorm(T1)
    Q[r, 2] <- 1 - pt(T1, min(n-1, m-1))
    Q[r, 3] <- 1 - pt(T2, n+m-2)
}
```

Control of Type I Error Rate

- What can you see about type I error rate for each test?

```
> apply(Q, 2, function(x) mean(x < 0.10))
[1] 0.1200 0.0847 0.1028
> apply(Q, 2, function(x) mean(x < 0.05))
[1] 0.0712 0.0320 0.0508
> apply(Q, 2, function(x) mean(x < 0.01))
[1] 0.0222 0.0030 0.0116
```

  - The Z-test has an inflation of type I error rate.
  - The T_1-test has too conservative type I error rate.
  - The T_2-test controls type I error rate well.

Statistical Power Example

- The following simulation studies compare the powers of T_1-test and T_2-test,

$$T_1 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}} \quad \text{vs.} \quad T_2 = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

  when testing

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x > \mu_y$$

- $X \sim N(\mu_x, \sigma_x)$ and $Y \sim N(\mu_y, \sigma_y)$
- In the first simulation,
  - $\mu_x = 1$, $\mu_y = 0$, $\sigma_x = \sigma_y = 1$
- In the second simulation,
  - $\mu_x = 3$, $\mu_y = 0$, $\sigma_x = 1$, $\sigma_y = 3$

```
set.seed(9876)

P <- matrix(0, nrep, 2)
mu_x <- 1; mu_y <- 0; sig_x <- sig_y <- 1
for (r in 1:nrep) {
    X <- rnorm(n, mean=mu_x, sd=sig_x)
    Y <- rnorm(m, mean=mu_y, sd=sig_y)
    MD <- mean(X)-mean(Y)
    VX <- var(X)
    VY <- var(Y)
    T1 <- MD/sqrt(VX/n+VY/m)
    Sp2 <- ((n-1)*VX+(m-1)*VY)/(n+m-2)
    T2 <- MD/sqrt(Sp2*(1/n+1/m))
    P[r,1] <- 1-pt(T1, min(n-1, m-1))
    P[r,2] <- 1-pt(T2, n+m-2)
}
apply(P, 2, function(x) mean(x < 0.05))
```

```
set.seed(1234)

P <- matrix(0, nrep, 2)
mu_x <- 3; mu_y <- 0; sig_x <- 1; sig_y <- 3
for (r in 1:nrep) {
    X <- rnorm(n, mean=mu_x, sd=sig_x)
    Y <- rnorm(m, mean=mu_y, sd=sig_y)
    MD <- mean(X)-mean(Y)
    VX <- var(X)
    VY <- var(Y)
    T1 <- MD/sqrt(VX/n+VY/m)
    Sp2 <- ((n-1)*VX+(m-1)*VY)/(n+m-2)
    T2 <- MD/sqrt(Sp2*(1/n+1/m))
    P[r,1] <- 1-pt(T1, min(n-1, m-1))
    P[r,2] <- 1-pt(T2, n+m-2)
}
apply(P, 2, function(x) mean(x < 0.05))
```

Duality between Confidence Intervals and Tests

- Suppose we carry out the following test regarding a single population mean at a significance level of α.

$$H_0 : \mu = \mu_0 \qquad \text{vs.} \qquad \mu \neq \mu_0$$

- Then, we can reject H_0 if and only if μ_0 does not belong to the $100(1-\alpha)$% confidence interval for μ.
- This is called a duality between confidence intervals and hypotheses tests.
- Duality is limited to only two side tests.
- If μ_0 belongs to a confidence interval, then it is a credible value for the population mean and hence we do not reject H_0
- Confidence regions (acceptance regions) $+$ Rejection regions $=$ 100%

```
set.seed(5678)
mu <- 0.3
alp <- c(0.01, 0.05, 0.1)
nrep <- 1e4
POW <- CP <- matrix(0, length(alp), 2)
n <- 100

for (i in 1:length(alp)) {
    X <- matrix(rnorm(n*nrep, mu), nrep, n)
    MX <- apply(X, 1, mean)
    SX <- apply(X, 1, sd)
    Z <- MX/(1/sqrt(n))
    T <- MX/(SX/sqrt(n))
    q0 <- qnorm(1-alp[i]/2)
    q1 <- qt(1-alp[i]/2, df=n-1)
    POW[i, 1] <- mean(abs(Z) > q0)
    POW[i, 2] <- mean(abs(T) > q1)
    CP[i, 1] <- 1 - mean(MX-q0/sqrt(n) < 0
                        & MX+q0/sqrt(n) > 0)
    CP[i, 2] <- 1 - mean(MX-q1*SX/sqrt(n) < 0
                        & MX+q1*SX/sqrt(n) > 0)
}

rownames(POW) <- rownames(CP) <- alp
colnames(POW) <- colnames(CP) <- c("Z-test", "T-test")
POW
CP
```

Permutation Test

- Permutation test is resampling-based test. It calculates p-values in a completely different way, working 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.

Permutation Test: Two-sample Test

- For the two-sample test for a population mean, we test $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$.
- Based on the observed data X_i for $i = 1, \ldots, n$ and Y_j for $j = 1, \ldots, m$, we can use the following test statistic,

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

  where $\bar{X}$ and $\bar{Y}$ are the sample means and $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ are the sample variances for X_i and Y_j, respectively.
- 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.

Permutation Test: Two-sample Test

- 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, 3, 4, 2, 2, 1, 2

- 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
  - The artificial Y set is 4, 2, 2, 1.

Permutation Test: Two-sample Test

- For the k-th permuted data set, a null test statistics T_k is

$$T_k = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}},$$

- The two-sided permutation p-value is then simply

$$\frac{1}{K} \sum_{k=1}^{K} I\left(|T_k| > |T_{\textsf{obs}}|\right)$$

  where K is the total number of permutations and $I(\cdot)$ is the indicator function.

- The following code gives the permutation test p-value for the null hypothesis that the population means for the two populations are equal.

```
set.seed(1234)

n <- 20; m <- 10
X <- rnorm(n, mean=0, sd=1)
Y <- rnorm(m, mean=2, sd=2)

## Calculate the test statistic for the original data
mx <- mean(X)
my <- mean(Y)
vx <- var(X)
vy <- var(Y)
T <- (mx - my) / sqrt(vx/n + vy/m)

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

```
## Get 10000 test statistics for permuted data.
TR <- array(0, 10000)
for (r in (1:10000)) {
    ## Generate a random permutation.
    ii <- sample(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 test stat for the reassigned data.
    mx <- mean(x);vx <- var(x)
    my <- mean(y);vy <- var(y)
    TR[r] <- (mx - my) / sqrt(vx/n + vy/m)
}
```

```
## A two-sided p-value.
pv2 <- mean(abs(TR) > abs(T))
pv2

## A one-sided left-tailed p-value
pvr <- mean(TR > T)
pvr

## A one-sided right-tailed p-value
pvr <- mean(TR < T)
pvr

## The null distribution
hist(TR, nclass=50, col="orange", freq=FALSE, main="")
abline(v=c(-T, T), lty=2, col=2)
```

Performance of Permutation Test

- The permutation test is an appealing idea, but we should check that it actually works.
- We investigate the type I error rate and power.
- The first simulation compute the type I error rate when the two populations being compared are normal with mean zero, but have different variances. The variances are generated independently from a standard exponential distribution.
- The second simulation compare the powers of a theoretical test and a permutation test when two populations have a different mean but the same variance. We compute both p-values of two sample T-test and permutation test for each simulation replication when the mean difference set as (0.5, 1, and 1.5) and the sample sizes are (10, 20, 30, and 50).

```
set.seed(1234)
n <- 10
m <- 10
nrep <- 1000

pv <- array(0, nrep)
for (j in 1:nrep) {
    V <- rexp(2)
    X <- rnorm(n, sd=sqrt(V[1]))
    Y <- rnorm(m, sd=sqrt(V[2]))

    mx <- mean(X); vx <- var(X)
    my <- mean(Y); vy <- var(Y)
    T <- (mx - my) / sqrt(vx/n + vy/m)

    Z <- c(X, Y)
```

```
   TR <- array(0, 1000)
   for (r in (1:1000)) {
       ii <- sample(m+n)
       x <- Z[ii[1:n]]
       y <- Z[ii[(n+1):(n+m)]]
       mx <- mean(x); vx <- var(x)
       my <- mean(y); vy <- var(y)
       TR[r] <- (mx - my) / sqrt(vx/n + vy/m)
    }
    pv[j] <- mean(abs(TR) > abs(T))
}

## Type I error rate for alpha=0.01, 0.05 and 0.1
c(mean(pv < 0.01), mean(pv < 0.05), mean(pv < 0.1))
```

```
set.seed(1111)
K <- 100
mu <- c(0.5, 1, 1.5)
nsamp <- c(10, 20, 30, 50)

for (i in 1:length(mu)) {
    d <- mu[i]
    for (j in 1:length(nsamp)) {
        n <- nsamp[j]
        pw1 <- pw2 <- NULL
        for (k in 1:K) {
            X <- rnorm(n, mean = d, sd = 2)
            Y <- rnorm(n, mean = 0, sd = 2)
            MX <- mean(X); VX <- var(Y)
            MY <- mean(Y); VY <- var(Y)
            Sp2 <- ((n-1)*VX + (n-1)*VY)/(2*n-2)
            TS <- (MX - MY)/sqrt(2*Sp2/n)
            pw1[k] <- ((1 - pt(TS, 2*n-2)) < 0.05)
            Z <- c(X, Y)
            TR <- NULL
            for (r in 1:1000) {
                ii <- sample(2*n)
                x <- Z[ii[1:n]]
                y <- Z[ii[(n+1):(2*n)]]
                mx <- mean(x); vx = var(x)
                my <- mean(y); vy = var(y)
                sp <- ((n-1)*vx + (n-1)*vy)/(2*n-2)
                TR[r] <- (mx - my)/sqrt(2*sp/n)
            }
            pw2[k] <- (mean(TR > TS) < 0.05)
        }
        if (i+j == 2) RE <- c(d, n, mean(pw1), mean(pw2))
        else RE <- rbind(RE, c(d, n, mean(pw1), mean(pw2)))
    }
}

colnames(RE) <- c(’mu’, ’sample’, ’t-test’, ’permutation’)
rownames(RE) <- seq(length(mu)*length(nsamp))
RE
```