

My Statistics Note

Contents

1	Sampling Distributions Related to the Normal Distribution	1
2	Hypothesis Testing	2
2.1	Testing Proportions	2
2.2	Testing Categorical Variables	5
3	Simple Linear Regression	7
3.1	An Example of Linear Relationship	7
3.2	The coefficients:	8
3.3	Confidence Intervals	8
3.4	The Multiple R^2	8
3.5	Prediction	9

1 Sampling Distributions Related to the Normal Distribution

Theorem 1.1. Let Y_1, Y_2, \dots, Y_n be a random sample of size n from a normal distribution with mean μ and variance σ^2 . Then

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

is normally distributed with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma_{\bar{Y}}^2 = \sigma^2/n$.

Theorem 1.2. Let Y_1, Y_2, \dots, Y_n be defined as in Theorem 1.1. Then $Z_i = \frac{Y_i - \mu}{\sigma}$ are independent, standard normal random variables, $i = 1, 2, \dots, n$, and

$$\sum_{i=1}^n Z^2 = \sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with n degrees of freedom (df).

A good estimator of σ^2 is the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Theorem 1.3. Let Y_1, Y_2, \dots, Y_n be a random sample from a normal distribution with mean μ and variance σ^2 . Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

has a χ^2 distribution with $(n-1)$ df. Also, \bar{Y} and S^2 are independent random variables.

Definition 1.1. Let Z be a standard normal random variable and let W be a χ^2 -distributed variable with ν df. Then if Z and W are independent,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

is said to have a t distribution with ν df.

If Y_1, \dots, Y_n constitute a random sample from a normal population with mean μ and variance σ^2 . Then,

$$Z = \sqrt{n}(\bar{Y} - \mu)/\sigma$$

has a standard normal distribution (Theorem 1.1).

$$W = (n-1)S^2/\sigma^2$$

has a χ^2 distribution with df $\nu = n-1$ (Theorem 1.3). And Z and W are independent. Therefore,

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}(\bar{Y} - \mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \sqrt{n} \left(\frac{\bar{Y} - \mu}{S} \right)$$

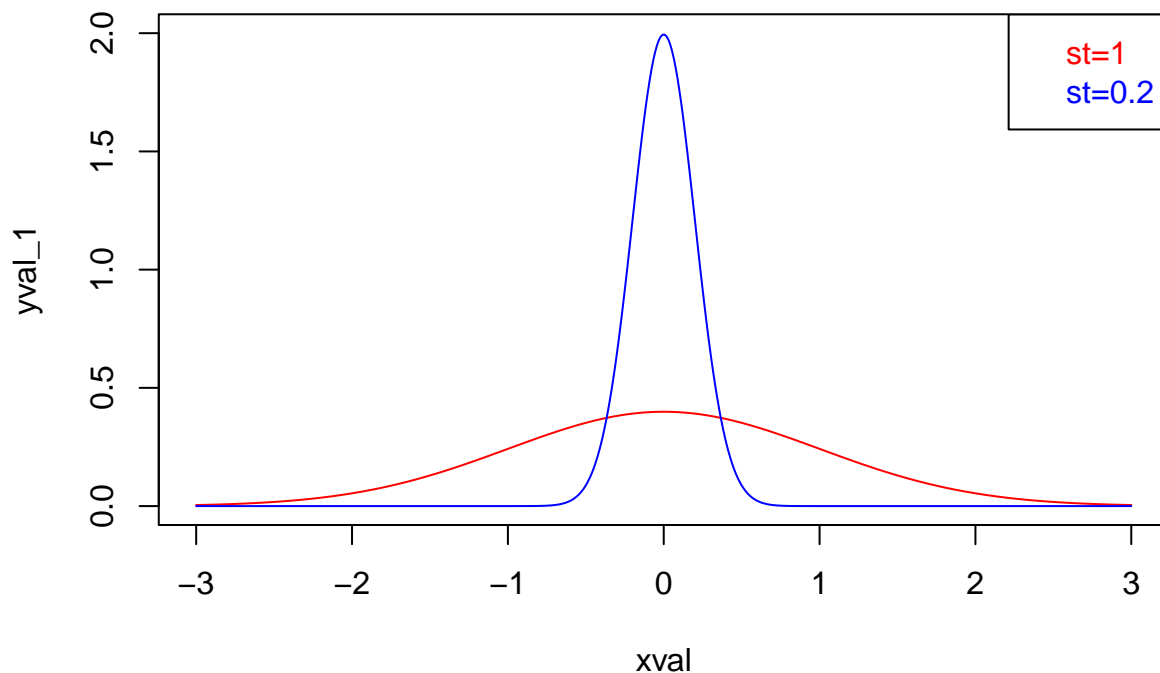
has a t distribution with $(n-1)$ df.

Definition 1.2. F distribution...

Graph of Gaussian distribution

```
xval <- seq(-3, 3, length.out = 500)
yval_1 <- dnorm(xval, mean=0, sd=1)
yval_0_2 <- dnorm(xval, mean=0, sd=0.2)
```

```
plot(xval, yval_1, type='l', ylim = c(0,2), col="red")
points(xval, yval_0_2, type='l', col="blue")
legend("topright", legend = c("st=1", "st=0.2"), text.col=c("red", "blue"))
```



?legend

2 Hypothesis Testing

2.1 Testing Proportions

2.1.1 Single Proportion

Consider n binary trials, in which the results are success (1) and failure (0). Denote by π the true proportion, \hat{p} is the sample estimate of π . A rule of thumb: $n\hat{p}$ and $n(1 - \hat{p})$ are both greater than 5.

One-Sample Z-Test

$$H_0 : \pi = \pi_0$$

$$H_A : \pi \neq \pi_0$$

Test statistic with the following

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$

We can assume $Z \sim N(0, 1)$.

Two Proportions

$$H_0 : \pi_2 - \pi_1 = 0$$

$$H_A : \pi_2 - \pi_1 > 0.$$

Let $\hat{p}_1 = x_1/n_1$, $\hat{p}_2 = x_2/n_2$, the test statistic is given by

$$Z = \frac{\hat{p}_2 - \hat{p}_1 - \pi_0}{\sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}},$$

where

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}.$$

We can treat $Z \sim N(0, 1)$.

Example 2.1. Consider the case:

$$H_0 : \pi = 0.9$$

$$H_A : \pi < 0.9$$

$n = 89$ and $x = 71$.

1. Compute the test statistic and the p -value and state your conclusion for the test using a significance level of $\alpha = 0.1$.
2. Using your estimated sample proportion, construct a two-sided 90 percent confidence interval for the true proportion of women who would recommend the skin cream.

```
# Q1
prop.test(71, 89, p=0.9, alternative="less", conf.level = 0.9, correct=FALSE)
```

```
##
## 1-sample proportions test without continuity correction
##
## data: 71 out of 89, null probability 0.9
```

```
## X-squared = 10.338, df = 1, p-value = 0.0006515
## alternative hypothesis: true p is less than 0.9
## 90 percent confidence interval:
## 0.0000000 0.8466949
## sample estimates:
##      p
## 0.7977528
```

```
n <- 89
p_hat <- 71/89
pi_0 <- 0.9
Z <- (p_hat - pi_0) / (sqrt(pi_0*(1-pi_0)/n))
# The p-value = pnorm(Z)
pnorm(Z)
```

```
## [1] 0.0006514802
```

p -value very small; less than 0.1. There is evidence to reject H_0 and conclude the true proportion of women who would recommend in samples of size 89 is less than 0.9.

```
# Q2
qnorm(c(0.05, 0.95)) * sqrt(p_hat * (1-p_hat)/n) + p_hat
```

```
## [1] 0.7277190 0.8677866
```

Example 2.2. Consider the case:

$$\begin{aligned}x_1 &= 97, & n_1 &= 445, \\x_2 &= 90, & n_2 &= 419,\end{aligned}$$

$$\begin{aligned}H_0 &: \pi_2 - \pi_1 = 0 \\H_A &: \pi_2 - \pi_1 \neq 0.\end{aligned}$$

```
prop.test(x=c(97, 90), n=c(445, 419), alternative = "two.sided", conf.level = 0.95, correct=FALSE)
```

```
##
## 2-sample test for equality of proportions without continuity correction
##
## data: c(97, 90) out of c(445, 419)
## X-squared = 0.012871, df = 1, p-value = 0.9097
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.05175429 0.05811507
## sample estimates:
##      prop 1      prop 2
## 0.2179775 0.2147971
```

```
x_1 <- 97
x_2 <- 90
n_1 <- 445
n_2 <- 419

p_hat_1 <- x_1 / n_1
p_hat_2 <- x_2 / n_2
p_star <- (x_1 + x_2) / (n_1 + n_2)
```

#The 95% confidence interval

```
qnorm(c(0.025, 1-0.025)) * sqrt(p_star*(1-p_star)*(1/n_1+1/n_2)) + p_hat_2 - p_hat_1
```

```
## [1] -0.05812427 0.05176349
```

Write a function Z.test that can perform a one- or two-sample Z-test

```
Z.test <- function (p1, n1, p2=NULL, n2=NULL, p0, alternative="two.sided", conf.level=0.95){
  if (!is.null(p2) & !is.null(n2)) {
    cat("Two-sample Z-test.\n")
    p.star <- (p1 * n1 + p2 * n2) / (n1 + n2)
    z <- (p1 - p2 - p0) / sqrt(p.star*(1-p.star)*(1/n1 + 1/n2))
    ci <- qnorm(c( (1 - conf.level) / 2, 1 - (1-conf.level) / 2)) * sqrt(p.star*(1-p.star)*(1/n1 + 1/n2))
    if (alternative == "two.sided") {
      if (z<0) {
        p <- 2 * pnorm(z)
      } else if (z>=0) {
        p = 2*(1-z)
      }
    } else if (alternative == "greater") {
      p <- 1 - pnorm(z)
    } else if (alternative == "less") {
      p <- pnorm(z)
    }
  } else if (!is.null(p2) | !is.null(n2)) {
    cat("One-sample Z-test.\n")
    z <- (p1 - p0) / sqrt(p0*(1-p0) / n1)
    if (alternative == "two.sided") {
      if (z<0) {
        p <- 2 * pnorm(z)
      } else if (z>=0) {
        p = 2*(1-z)
      }
    } else if (alternative == "greater") {
      p <- 1 - pnorm(z)
    } else if (alternative == "less") {
      p <- pnorm(z)
    }
  }

  ci <- qnorm(c( (1 - conf.level) / 2, 1 - (1-conf.level) / 2)) * sqrt(p0*(1-p0) / n1) + p1
}
return(list(Z=z,P=p,CI=ci))
}
```

```
x1 <- 180
n1 <- 233
p.hat1 <- x1/n1
x2 <- 175
n2 <- 197
p.hat2 <- x2/n2
```

```
Z.test(p.hat2,n2,NULL,n1,p0=0.2,conf.level=0.95, alternative = "less") # ...or you could flip the order
```

```
## One-sample Z-test.
```

```
## $Z
```

```
## [1] 24.15275
```

```
##
## $P
## [1] 1
##
## $CI
## [1] 0.8324682 0.9441815
```

2.2 Testing Categorical Variables

2.2.1 Single Categorical Variable

Consider the following example. Suppose a researcher in sociology is interested in the dispersion of rates of facial hair in men of his local city and whether they are uniformly represented in the male population. He defines a categorical variable with three levels: clean shaven (1), beard only or moustache only (2), and beard and moustache (3). He collects data on 53 randomly selected men and finds the following outcomes:

```
hairy <- c(2,3,2,3,2,1,3,3,2,2,3,2,2,2,3,3,3,2,3,2,2,2,1,3,2,2,2,1,2,2,3,
          2,2,2,2,1,2,1,1,1,2,2,2,3,1,2,1,2,1,2,1,3,3)
```

The research question asks whether the proportions in each category are equally represented. Let π_1 , π_2 , and π_3 represent the true proportion of men in the city who fall into groups 1, 2, and 3, respectively. You therefore seek to test these hypotheses:

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$$

$$H_A : H_0 \text{ is incorrect}$$

For this test, use a standard significance level of 0.05.

Calculation: Chi-Squared Test of Distribution

The quantities of interest are the proportion of n observations in each of k categories, π_1, \dots, π_k , for a single mutually exclusive and exhaustive categorical variable. The null hypothesis defines hypothesized null values for each proportion; label these respectively as $\pi_{0(1)}, \dots, \pi_{0(k)}$. The test statistic χ^2 is given as

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed count and E_i is the expected count in the i th category, $i = 1, \dots, k$.

1. O_i are obtained directly from the raw data.
2. $E_i = n\pi_{0(i)}$ are merely the product of the overall sample size n with the respective null proportion for each category.

The result of χ^2 follows a χ^2 -distribution with $\nu = k - 1$ degree of freedom.

```
n <- length(hairy)
hairy.tab <- table(hairy)
hairy.tab / n

## hairy
##      1      2      3
## 0.2075472 0.5283019 0.2641509

expected <- 1/3 * n
hairy.matrix <- cbind(1:3, hairy.tab, expected,
                     (hairy.tab-expected)^2/expected)
dimnames(hairy.matrix) <- list(c("clean", "beard OR mous."),
```

```

                                "beard AND mous."),
                                c("i", "Oi", "Ei", "(Oi-Ei)^2/Ei"))
hairy.matrix

```

```

##           i Oi           Ei (Oi-Ei)^2/Ei
## clean      1 11 17.66667    2.5157233
## beard OR mous. 2 28 17.66667    6.0440252
## beard AND mous. 3 14 17.66667    0.7610063

```

```

X2 <- sum(hairy.matrix[,4])
X2

```

```

## [1] 9.320755

```

```

1-pchisq(X2, df=2)

```

```

## [1] 0.009462891

```

```

chisq.test(x=hairy.tab)

```

```

##
## Chi-squared test for given probabilities
##
## data: hairy.tab
## X-squared = 9.3208, df = 2, p-value = 0.009463

```

This small p -value provides evidence to suggest that the true frequencies in the defined categories of male facial hair are not uniformly distributed in a $1/3, 1/3, 1/3$ fashion.

```

chisq.test(x=hairy.tab, p=c(0.25, 0.5, 0.25))

```

```

##
## Chi-squared test for given probabilities
##
## data: hairy.tab
## X-squared = 0.50943, df = 2, p-value = 0.7751

```

The high p -value suggests there is no evidence to reject H_0 in this scenario. In other words, there is no evidence to suggest that the proportions hypothesized in H_0 are incorrect.

3 Simple Linear Regression

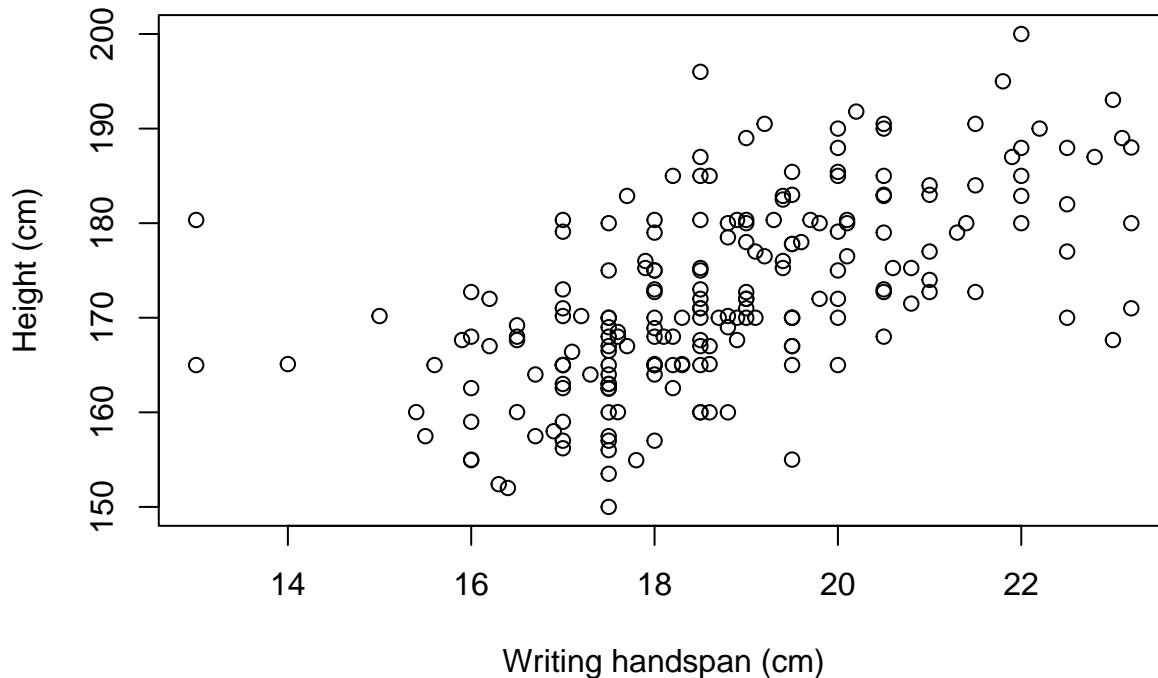
3.1 An Example of Linear Relationship

Consider the example survey in MASS.

```

library("MASS")
plot(survey$Height~survey$Wr.Hnd, xlab="Writing handspan (cm)", ylab="Height (cm)")

```



Consider the simple variable linear model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where x is the variable `survey$Wr.Hnd`, \hat{y} is the variable `survey$Height`.

```
survfit <- lm(Height~Wr.Hnd, data=survey)
summary(survfit)
```

```
##
## Call:
## lm(formula = Height ~ Wr.Hnd, data = survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7276  -5.0706  -0.8269   4.9473  25.8704
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  113.9536     5.4416   20.94  <2e-16 ***
## Wr.Hnd        3.1166     0.2888   10.79  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.909 on 206 degrees of freedom
## (29 observations deleted due to missingness)
## Multiple R-squared:  0.3612, Adjusted R-squared:  0.3581
## F-statistic: 116.5 on 1 and 206 DF, p-value: < 2.2e-16
```

3.2 The coefficients:

As is shown,

$$\hat{\beta}_0 = 113.9536,$$

$$\hat{\beta}_1 = 3.1166.$$

The parameters follow t -distributions with degrees of freedom $(n - 2)$. The standardized t value and p -value are reported for each parameter:

$$113.9536/5.4416 = 20.94,$$

$$3.1166/0.2888 = 10.79.$$

These represent the results of a two-tailed hypothesis test formally defined as

$$H_0 : \beta_j = 0$$

$$H_A : \beta_j \neq 0.$$

3.3 Confidence Intervals

```
confint(survfit, level=0.95)
```

```
##                2.5 %      97.5 %
## (Intercept) 103.225178 124.682069
## Wr.Hnd      2.547273   3.685961
```

3.4 The Multiple R^2

The multiple R^2 tells that about 36.1% of the variation in the student heights can be attributed to handspan.

```
rho_xy <- cor(survey$Height, survey$Wr.Hnd, use="complete.obs")
rho_xy^2
```

```
## [1] 0.3611901
```

3.5 Prediction

3.5.1 Confidence Intervals for Mean Heights

```
xvals <- data.frame(Wr.Hnd=c(14.5, 24))
mypred.ci <- predict(survfit, newdata = xvals, interval="confidence", level=0.95)
mypred.ci
```

```
##      fit      lwr      upr
## 1 159.1446 156.4956 161.7936
## 2 188.7524 185.5726 191.9323
```

3.5.2 Prediction Intervals for Individual Observations

```
mypred.pi <- predict(survfit, newdata = xvals, interval="prediction", level=0.95)
mypred.pi
```

```
##      fit      lwr      upr
## 1 159.1446 143.3286 174.9605
## 2 188.7524 172.8390 204.6659
```