Tutorial for STA2002

Kun HUANG(SDS)

2020-10-27

Contents

Prerequisites

Probability and Statistics I(STA2001) is the prerequisite, which mainly includes the following contents,

- Some usual distributions, like Binomial, Poisson, Normal, Exponential, Gamma, and Chi-square distributions (Relationships among some univariate distributions(Song 2005));
- Basic terminologies, e.g.,independence, expectation, variation, correlation (coefficient), Bayes, and etc;
- Large number theorem, like Central Limit Theorem(CLT).

Chapter 1

Tutorial 1

1.1 Q1

- Moment-generating function $M(t)$ of a random variable X defined in D that has a density function $f(x)$.

$$M(t) = \mathbb{E}(e^{tx}) = \int_D e^{tx} f(x) dx \tag{1.1}$$

$$\mathbb{E}(X^s) = M^{(s)}(0) \tag{1.2}$$

- Relationship between $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, independent.
- How to derive a quantity following t distribution from a norm population.

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \tag{1.3}$$

- The t distribution is symmetric, i.e., $t_q(n) = -t_{1-q}(n), q \in (0,1)$. For example,

```
qt(0.025, 8, lower.tail = F)
```

```
## [1] 2.306004
```

```
-qt(1 - 0.025, 8, lower.tail = F)
```

```
## [1] 2.306004
```

- Properties of F distribution: $F_{0.95}(9, 24) = \frac{1}{F_{0.05}(24,9)}$

1.2 Q2

- Standardize a norm distribution $X \in \mathcal{N}(\mu, \sigma^2)$, i.e., $\frac{X-\mu}{\sigma} \in \mathcal{N}(0,1)$.
- The distribution of $\bar{X}$ and S^2.

1.3 Q3

- Central Limit Theorem(CLT)

Theorem 1.1. ***(Central Limit Theorem)*** *Let $X_1, \ldots, X_n$ be independent, identically distributed (i.i.d.) random variables with finite expectation μ, and positive, finite variance σ^2, and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{L} N(0,1) \text{ as } n \to \infty.$$

- The relationship between Binomial distribution $b(n,p)$ and Poisson distribution Pois(λ): $\infty > np = \lambda, n \to \infty$
- Aware the power of CLT.

Chapter 2

Tutorial 2

2.1 Q1

- Derive moments from a given pdf $f(x)$. $EX = \int xf(x)dx, EX^2 = \int x^2 f(x)dx$.
- Derive variance from the first and second moments,i.e., $Var(X) = EX^2 - E^2X$.
- $E(aX + bY + c) = aEX + bEY + c$, $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y)$. The latter needs X and Y are independent.
- CLT approximation.

2.2 Q2

Definition 2.1 (Poisson Process)**.** Let $N(t)$ be the number of events happens during the time interval $[0, t]$, if $N(t)$ satisfies the following:

- N(0) = 0;
- has independent increments, and
- $\forall\tau > 0$, $P(N(t + \tau) - N(t) = n) = \frac{(\lambda\tau)^n}{n!}e^{-\lambda\tau}$

we call $\{N(t), t \geq 0\}$ is a Poisson process with rate λ.

- Let $(W_n > t)$ be the $n - th$ random event happens after time t, then $W_n \sim Gamma(n, \lambda)$. In fact, $Gamma(n, \lambda)$ can be seen as the time to be waited until the $n - th$ event.

2.3 Q3

Proposition 2.1. *Suppose the random variable X has a pdf $f(x)$, let $Y = T(X)$, where $T : \mathbb{R} \to \mathbb{R}$ is an invertible transformation. Then the pdf $g(y)$ of Y is*

$$g(y) = f(T^{-1}(y))\frac{dT^{-1}(y)}{dy}$$

For example, suppose $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = aX + b \sim \mathcal{N}(a\mu + b, (a\sigma)^2)$.

2.4 Q4

Theorem 2.1 (Chebyshev's Inequality)**.** *Let X be a random variable with finite mean μ and variance $\sigma^2 > 0$. Then $\forall k > 0$,*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Additionally, let $k\sigma = \varepsilon$, the above becomes,

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Chapter 3

Tutorial 3

3.1 Method of moment estimator(MME)

Suppose that the problem is to estimate k unknown parameters $\boldsymbol{\theta} := (\theta_1, \theta_2, ..., \theta_k)^T$ characterizing the distribution $f_X(x; \boldsymbol{\theta})$ of the random variable X. Suppose the first k moments of the true distribution can be expressed by the function of $\boldsymbol{\theta}$, i.e.,

$$\mu_1 \equiv \mathrm{E}[W] = g_1(\theta_1, \theta_2, \ldots, \theta_k) \tag{3.1}$$

$$\mu_2 \equiv \mathrm{E}\left[W^2\right] = g_2(\theta_1, \theta_2, \ldots, \theta_k) \tag{3.2}$$

$$\vdots \tag{3.3}$$

$$\mu_k \equiv \mathrm{E}\left[W^k\right] = g_k(\theta_1, \theta_2, \ldots, \theta_k) \tag{3.4}$$

Suppose a sample of size n is drawn, having the values of $x_1, x_2, ..., x_n$, let

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j, j = 1, 2, ..., k \tag{3.5}$$

Solve the above k equations, we derive the method of moment estimator of $\boldsymbol{\theta}$.

3.2 Maximum likelihood estimator(MLE).

Suppose we have a sample of size n, $X_1, ..., X_n$ i.i.d drawn from a population distribution $f_X(x; \boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)^T$. Define the likelihood function to be

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

The log-likelihood function is defined by $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is determined to maximize $L(\boldsymbol{\theta})$, i.e.,

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) \tag{3.6}$$

Chapter 4

Confidence Interval

Definition 4.1 (Confidence Interval)**.** Given a sample $X_1, X_2, ..., X_n$ of the population $X \sim f(x;\theta)$ and $\alpha \in [0,1]$, a $(1-\alpha)$ confidence interval $(a(X_1, X_2, ..., X_n), b(X_1, X_2, ..., X_n))$ for the parameter θ is defined such that,

$$P\left[a(X_1, X_2, ..., X_n) < \theta < b(X_1, X_2, ..., X_n)\right] = 1-\alpha \tag{4.1}$$

Interpretation and misunderstanding

4.1 Q1

Definition 4.2 (t-distribution)**.** Suppose $X \sim N(0,1)$, $U \sim \chi^2(n)$, and X are independent from Y, then $\frac{X}{\sqrt{U/n}}$ has a (student) t distribution with n degrees of freedom, i.e.,

$$\frac{X}{\sqrt{U/n}} \sim t(n)$$

Confidence Interval of normal population $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2), i = 1, 2, ..., n$. We have

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \tag{4.2}$$

It can be proved that $\bar{X}$ and S^2 are independent. Then,

$$\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \tag{4.3}$$

We call such a method the pivotal approach. A pivotal quantity or pivot is a function of observations and unobservable parameters such that the function's probability distribution does not depend on the unknown parameters. For example, $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$ is a pivot. From (4.3), we derive the $(1-\alpha)$ confidence interval for the mean μ when σ^2 is unknown, i.e.,

$$\bar{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} \tag{4.4}$$

```
qt(0.05/2, 8, lower.tail = F)
```

```
## [1] 2.306004
```

```
qnorm(0.01/2, lower.tail = F)
```

```
## [1] 2.575829
```

4.2 Q2

Theorem 4.1 (Welch's t-interval). *Let $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, ..., Y_m \overset{i.i.d.}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ be independent random variables. Then an approximate $(1-\alpha)$ C.I. for $\mu_X - \mu_Y$ is*

$$\bar{X} - \bar{Y} \pm t_{\alpha/2}(r)\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

where

$$r = \left\lfloor \frac{\left(\frac{S_X^2}{n} + \frac{S_X^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2} \right\rfloor$$

Let $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, ..., Y_m \overset{i.i.d.}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ be independent random variables. We have the following,

$$\bar{X} \sim \mathcal{N}(\mu_X, \frac{\sigma_X^2}{n}), \quad \bar{Y} \sim \mathcal{N}(\mu_Y, \frac{\sigma_Y^2}{n}) \tag{4.5}$$

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1), \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(m-1) \tag{4.6}$$

The two samples are independent, hence,

$$\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}) \tag{4.7}$$

- $\sigma_X = \sigma_Y = \sigma$ and σ is known, then,

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim \mathcal{N}(0,1) \tag{4.8}$$

- $\sigma_X = \sigma_Y = \sigma$ and σ is unknown, then,

$$\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} \sim \chi^2(n+m-2) \tag{4.9}$$

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)/\left(\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}\right)}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{\sigma^2(n+m-2)}}} \sim t(n+m-2) \tag{4.10}$$

- $\sigma_X \neq \sigma_Y$ and they are both unknown, use Welch's t-interval or CLT approximation.

- $m = n$, then $Z_i = X_i - Y_i \sim \mathcal{N}(\mu_X - \mu_y, \sigma_Z)$ since $(X_i, Y_i)^T \sim \mathcal{N}\left((\mu_X, \mu_Y)^T, \Sigma\right)$. Then the same technique in Q1 can be used.

```
qt(0.05 / 2, 8, lower.tail = F)
```

```
## [1] 2.306004
```

4.3 Q3

If $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$ are independent, then

$$\frac{X/n}{Y/m} \sim F(n, m)$$

Therefore, with samples from two independent normal population, i.e., let $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, ..., Y_m \overset{i.i.d.}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ be independent, we have a pivot

$$\frac{\frac{(n-1)S_X^2}{\sigma_X^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2}/(m-1)} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1) \tag{4.11}$$

```
alpha <- 0.02
qf(alpha / 2, 12, 8, lower.tail = F)
```

```
## [1] 5.666719
```

```
qf(alpha / 2, 8, 12, lower.tail = F)
```

```
## [1] 4.499365
```

$$F_{1-\alpha/2}(r_1, r_2) = \frac{1}{F_{\alpha/2}(r_2, r_1)} \tag{4.12}$$

4.4 Q4

According to the central limit theorem(CLT), we have an approximate pivot

$$\frac{\bar{X} - EX}{\sqrt{VarX}} \to \mathcal{N}(0, 1) \tag{4.13}$$

```
qnorm(0.05 / 2, lower.tail = F)
```

```
## [1] 1.959964
```

4.5 Solutions

4.5.1 Q1

```
x <- c(21.5, 18.95, 18.55, 19.4, 19.15, 22.35, 22.9, 22.2, 23.1)
t.test(x, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 33.738, df = 8, p-value = 6.506e-10
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.47149 22.32851
## sample estimates:
## mean of x
##      20.9
```

```
n <- qnorm(0.1/2, lower.tail = F)^2 * var(x) / (0.5)^2
print(n)
```

```
## [1] 37.37708
```

4.5.2 Q2

```
x <- c(1612, 1352, 1456, 1222, 1560, 1456, 1924)
y <- c(1082, 1300, 1092, 1040, 910, 1248, 1092, 1040, 1092, 1288)
t.test(x,y, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 4.235, df = 8.5995, p-value = 0.002427
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  181.7191 604.9095
## sample estimates:
## mean of x mean of y
##  1511.714  1118.400
```

Note that R use $t_{\alpha/2}(8.6)$, so the result of C.I. is different from what we use where the df=8 in t distribution. The pdf of t distribution is

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{4.14}$$

where ν is the degree of freedom.

4.5.3 Q3

```
r1 <- 9 - 1
r2 <- 13 - 1
sx <- 128.41 / 12
sy <- 36.72 / 8
alpha <- 0.02
ci2 <- sx / sy * c(qf(1 - alpha / 2, r1, r2, lower.tail = F), qf(alpha / 2, r1, r2, lower.tail =
ci <- sqrt(ci2)
print(ci2)
```

```
## [1]  0.4114085 10.4895333
```

```
print(ci)
```

```
## [1] 0.6414113 3.2387549
```

4.5.4 Q4

$$\hat{p}_1 \pm z_{0.05/2}\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1}}$$

```
n1 <- 194
n2 <- 162
y1 <- 28
y2 <- 11
p1 <- y1 / n1
s1 <- sqrt(n1 * p1 * (1 - p1)) / n1
p1 + c(-1, 1) * qnorm(0.05/2, lower.tail = F) * s1
```

```
## [1] 0.0948785 0.1937813
```

$$z_{\alpha/2}\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n}} = \varepsilon$$

```
alpha <- 0.1
ep <- 0.04
qnorm(alpha / 2, lower.tail = F)^2 * p1 * (1 - p1) / ep^2
```

```
## [1] 208.8321
```

$$\left(\hat{p}_1 - \hat{p}_2\right) - z_{0.05}\sqrt{\frac{\hat{p}_1\left(1-\hat{\rho}_1\right)}{n_1} + \frac{\hat{\rho}_2\left(1-\hat{p}_2\right)}{n_2}}$$

```
p2 <- y2 / n2
p1 - p2 - qnorm(0.05, lower.tail = F) * sqrt(p1 * (1 - p1) / n1 + p2 * (1 -  p2) / n2)
```

```
## [1] 0.02370925
```

Chapter 5

Simple Linear Regression

Consider a simple linear regression model,

$$Y = \alpha + \beta(X - \bar{X}) + \varepsilon, \tag{5.1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Given X is not random, we have,

$$Y \sim \mathcal{N}(\alpha + \beta(X - \bar{X}), \sigma^2) \tag{5.2}$$

5.1 Fitting a Simple Linear Regression Model

Suppose we a series of samples $(x_i, y_i), i = 1, 2, ..., n$ and we want to fit a simple linear regression which has the form of (5.1). Then the fitted $(\hat{\alpha}, \hat{\beta})$ should minimize the residual, i.e.,

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha, \beta \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} (y_i - \alpha - \beta(x_i - \bar{x}))^2 \tag{5.3}$$

Solving (5.3), we derive

$$\hat{\alpha} = \bar{y}, \hat{\beta} = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{5.4}$$

Noting that $y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i \sim \mathcal{N}(\alpha + \beta(x_i - \bar{x}), \sigma^2)$, we have

$$\hat{\alpha} = \bar{y} \sim \mathcal{N}(\alpha, \frac{\sigma^2}{n}) \tag{5.5}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}) \tag{5.6}$$

The MLE for σ is

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{\alpha} - \hat{\beta}\left(x_i - \bar{x}\right)\right]^2 \tag{5.7}$$

5.2 A Toy Example

```r
# Simulated data
#' y = 4 + 3x + \epsilon
x <- runif(20, min = 5, max = 20)
y <- 4 + 3 * x + rnorm(20)
slr <- lm(y ~ x)
summary(slr)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25867 -0.65772 -0.03311  0.51021  1.96094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.68150    0.69319   5.311 4.76e-05 ***
## x            3.01429    0.05077  59.367  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.093 on 18 degrees of freedom
## Multiple R-squared:  0.9949,Adjusted R-squared:  0.9946
## F-statistic:  3524 on 1 and 18 DF,  p-value: < 2.2e-16
```

```r
names(slr)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```r
fitted(slr)
```

```
##        1        2        3        4        5        6        7        8
## 52.94912 34.45901 23.39097 26.15008 43.89705 63.73809 59.54406 24.01895
##        9       10       11       12       13       14       15       16
## 52.71892 61.02429 19.08463 50.77696 55.04924 33.73152 58.23259 53.07383
##       17       18       19       20
## 19.04014 38.87171 41.55155 32.51280
```

Chapter 6

Hypothesis Testing

6.1 Summary of Hypothesis Testing by Normal Population

Let samples $X_1, X_2, ..., X_n$ draw from a normal population $\mathcal{N}(\mu_X, \sigma_X^2)$, then,

	Distribution Under H_0	Critical Region
$H_0 : \mu = \mu_0$, σ known	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$	$H_1 : \mu > \mu_0, \quad \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$ $H_1 :$
$H_0 : \mu = \mu_0$, σ unknown	$\frac{\sqrt{n}(\bar{X}-\mu_0)/\sigma}{\sqrt{\frac{(n-1)S_X^2}{\sigma^2}/(n-1)}}$ $= \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} \sim t(n-1)$	$H_1 : \mu > \mu_0, \quad \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} \geq t_\alpha(n-1)$<br
$H_0 : \sigma^2 = \sigma_0^2$	$\frac{(n-1)S_X^2}{\sigma_0^2} \sim \chi^2(n-1)$	$H_1 : \sigma^2 > \sigma_0^2, \quad \frac{(n-1)S_X^2}{\sigma_0^2} \geq \chi_\alpha(n-1)$<

Null Hypothesis	Distribution Under H_0	Critical Region
$H_0 : \mu = \mu_0$, σ known	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$	$H_1 : \mu > \mu_0, \quad \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geq z_\alpha$ $H_1 : \mu < \mu_0, \quad \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leq z_{1-\alpha} = -z_\alpha$ $H_1 : \mu \neq \mu_0, \quad \lvert\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\rvert \geq z_{\alpha/2}$
$H_0 : \mu = \mu_0$, σ unknown	$\frac{\sqrt{n}(\bar{X}-\mu_0)/\sigma}{\sqrt{\frac{(n-1)S_X^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} \sim t(n-1)$	$H_1 : \mu > \mu_0, \quad \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} \geq t_\alpha(n-1)$ $H_1 : \mu < \mu_0, \quad \frac{\bar{X}-\mu_0}{S_X/\sqrt{n}} \leq t_{1-\alpha}(n-1)$ $H_1 : \mu \neq \mu_0, \quad \lvert\frac{\bar{X}-\mu_0}{S_X/\sqrt{n}}\rvert \geq t_{\alpha/2}$

Null Hypothesis	Distribution Under H_0	Critical Region
$H_0: \sigma^2 = \sigma_0^2$	$\frac{(n-1)S_X^2}{\sigma_0^2} \sim \chi^2(n-1)$	$H_1: \sigma^2 > \sigma_0^2, \quad \frac{(n-1)S_X^2}{\sigma_0^2} \geq \chi_\alpha(n-1)$ $H_1: \sigma^2 < \sigma_0^2, \quad \frac{(n-1)S_X^2}{\sigma_0^2} \leq \chi_{1-\alpha}(n-1) H_1: \sigma^2 \neq \sigma_0^2, \quad \frac{(n-1)S_X^2}{\sigma_0^2} \geq \chi_{\alpha/2}(n-1)$ or $\frac{(n-1)S_X^2}{\sigma_0^2} \leq \chi_{1-\alpha/2}(n-1)$

Let samples $X_1, X_2, ..., X_n$ draw from a normal population $\mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, ..., Y_m$ from another normal population $\mathcal{N}(\mu_Y, \sigma_Y^2)$.

Null Hypothesis	Distribution Under H_0
$H_0: \mu_X = \mu_Y$, $\sigma_X = \sigma_Y = \sigma$, known	$Z_1 = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1)$
$H_0: \mu_X = \mu_Y$, $\sigma_X = \sigma_Y = \sigma$, unknown	$T_1 = \frac{(\bar{X}-\bar{Y})/\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{\sigma^2}/(n+m-2)}}$ $= \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t($
$H_0: \mu_X = \mu_Y$, $\sigma_X \neq \sigma_Y$, unknown	$T_2 = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim t(r)$, $r = \left\lfloor \frac{\left(\frac{S_X^2}{n}+\frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2+\frac{1}{m-1}\left(\frac{S_Y^2}{m}\right.} \right.$
$H_0: \mu_X = \mu_Y$, $m = n$	$D_i := X_i - Y_i$ $\sim \mathcal{N}(\mu_X - \mu_Y, \sigma_Z^2)$ transform
$H_0: \sigma_X^2 = \sigma_Y^2$	$F = \frac{\frac{(n-1)S_X^2}{\sigma_X^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2}/(m-1)} = \frac{S_X^2}{S_Y^2}$ $\sim F(n-1, m-1)$

Null Hypothesis	Distribution Under H_0	Critical Region
$H_0: \mu_X = \mu_Y$, $\sigma_X = \sigma_Y = \sigma$, known	$Z_1 = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1)$	$H_1: \mu_X > \mu_Y, \quad Z_1 \geq z_\alpha H_1: \mu_X < \mu_Y, \quad Z_1 \leq z_{1-\alpha} = -z_\alpha H_1: \mu_X \neq \mu_Y, \quad \lvert Z_1 \rvert \geq z_{\alpha/2}$
$H_0: \mu_X = \mu_Y$, $\sigma_X = \sigma_Y = \sigma$, unknown	$T_1 = \frac{(\bar{X}-\bar{Y})/\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{\sigma^2}/(n+m-2)}} = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n+m-2)$	$H_1: \mu_X > \mu_Y, \quad T_1 \geq t_\alpha(n+m-2) H_1: \mu_X < \mu_Y, \quad T_1 \leq -t_\alpha(n+m-2) H_1: \mu_X \neq \mu_Y, \quad \lvert T_1 \rvert \geq t_{\alpha/2}(n+m-2)$

Null Hypothesis	Distribution Under H_0	Critical Region
$H_0 : \mu_X = \mu_Y, \sigma_X \neq \sigma_Y$, unknown	$T_2 = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim t(r)$, $r = \left\lfloor \frac{\left(\frac{S_X^2}{n}+\frac{S_Y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2+\frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2} \right\rfloor$	$H_1 : \mu_X > \mu_Y, \quad T_2 \geq t_\alpha(r) H_1 : \mu_X < \mu_Y, \quad T_2 \leq -t_\alpha(r) H_1 : \mu_X \neq \mu_Y, \quad \lvert T_2 \rvert \geq t_{\alpha/2}(r)$
$H_0 : \mu_X = \mu_Y, m = n$	$D_i := X_i - Y_i \sim \mathcal{N}(\mu_X - \mu_Y, \sigma_Z^2)$ transform it into the one sample situation with σ_Z^2 unknown	
$H_0 : \sigma_X^2 = \sigma_Y^2$	$F = \frac{\frac{(n-1)S_X^2}{\sigma_X^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2}/(m-1)} = \frac{S_X^2}{S_Y^2} \sim F(n-1, m-1)$	$H_1 : \sigma_X^2 > \sigma_Y^2, \quad F \geq F_\alpha(n-1, m-1) H_1 : \sigma_X^2 < \sigma_Y^2, \quad F \leq F_{1-\alpha}(n-1, m-1) H_1 : \sigma_X^{/} = \sigma_Y^2, \quad F \geq F_{\alpha/2}(n-1, m-1)$ or $F \leq F_{1-\alpha/2}(n-1, m-1)$

We next consider the situation of testing proportion. Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p_X)$ drawn from a specific event and $Y_i \overset{i.i.d.}{\sim} Bernoulli(p_Y)$. We want to infer p_X and the relationship between p_X and p_Y. Let $Z = \sum\limits_{i=1}^{n} X_i \sim Bin(n, p)$. $\hat{p} := \frac{Z}{n}$ is an unbiased estimator for p. According to the central limit theorem(CLT), we have,

$$\hat{p} \to \mathcal{N}(p, \frac{p(1-p)}{n}), n \to \infty \tag{6.1}$$

Under $H_0 : p_X = p_Y = p$, $\hat{p}_{XY} := \frac{\sum\limits_{i=1}^{n} X_i + \sum\limits_{i=1}^{m} Y_i}{n+m}$ is an unbiased estimator of p. Since,

$$\mathbb{E}(\hat{p}_{XY}) = \frac{\sum\limits_{i=1}^{n} \mathbb{E}X_i + \sum\limits_{i=1}^{m} \mathbb{E}Y_i}{n+m} = \frac{np + mp}{m+n} = p \tag{6.2}$$

Null Hypothesis	Distribution Under H_0	Criticall Region
$H_0 : p = p_0$	$Z_p = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{approx}{\sim} N(0, 1)$	$H_1 : p > p_0, \quad Z_p \geq$
$H_0 : p_X = p_Y$	$Z_{XY} = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\frac{\hat{p}_{XY}(1-\hat{p}_{XY})}{n}+\frac{\hat{p}_{XY}(1-\hat{p}_{XY})}{m}}}$ $\overset{apprrox}{\sim} \mathcal{N}(0, 1), \hat{p}_{XY} := \frac{\sum\limits_{i=1}^{n} X_i + \sum\limits_{i=1}^{m} Y_i}{n+m}$	$H_1 : p_X > p_Y, \quad Z$

Null Hypothesis	Distribution Under H_0	Criticall Region
$H_0: p = p_0$	$Z_p = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{approx}{\sim} N(0,1)$	$H_1: p > p_0, \quad Z_p \geq z_\alpha H_1: p < p_0, \quad Z_p \leq z_{1-\alpha} H_1: p \neq p_0, \quad \|Z_p\| \geq z_{\alpha/2}$
$H_0: p_X = p_Y$	$Z_{XY} = \frac{\hat{p}_X-\hat{p}_Y}{\sqrt{\frac{\hat{p}_{XY}(1-\hat{p}_{XY})}{n}+\frac{\hat{p}_{XY}(1-\hat{p}_{XY})}{m}}} \overset{apprrox}{\sim} \mathcal{N}(0,1), \hat{p}_{XY} := \frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i}{n+m}$	$H_1: p_X > p_Y, \quad Z_{XY} \geq z_\alpha H_1: p_X < p_Y, \quad Z_{XY} \leq z_{1-\alpha} H_1: p_X \neq p_Y, \quad \|Z_{XY}\| \geq z_{\alpha/2}$

6.2 Exercise

Exercise 6.1. To measure air pollution in a home, let X and Y equal the amount of suspended particulate matter (in g/m3) measured during a 24-hour period in a home in which there is no smoker and a home in which there is a smoker, respectively. We shall test the null hypothesis $H_0: \sigma_X^2/\sigma_Y^2 = 1$ against the one-sided alternative hypothesis $H_1: \sigma_X^2/\sigma_Y^2 > 1$. Suppose both samples are drawn from normal distribution.

1. If a random sample of size $n = 9$ yielded $\bar{x} = 93$ and $S_x = 12.9$ while a random sample of size $m = 11$ yielded $y = 132$ and $S_y = 7.1$, define a critical region and give your conclusion if $\alpha = 0.05$.
2. Now test $H_0: \mu_X = \mu_Y$ against $H_1: \mu_X < \mu_Y$ if $\alpha = 0.05$. $t_{0.05}(11) = 1.796$

Solutions:

1. To test $H_0: \sigma_X^2 = \sigma_Y^2$ against $H_1: \sigma_X^2 > \sigma_Y^2$ under normal populations.

$$F = \frac{S_x^2}{S_y^2} = \frac{12.9^2}{7.1^2} = 3.30 > 3.07 = F_{0.05}(8,10) \tag{6.3}$$

So we reject H_0 and conclude that $\sigma_X^2 \neq \sigma_Y^2$.

2. To test $H_0: \mu_X = \mu_Y$ against $H_1: \mu_X < \mu_Y$ under normal populations with variance not being equal.

$$r = \left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}\right] = \left[\frac{\left(\frac{12.9^2}{9}+\frac{7.1^2}{13}\right)^2}{\frac{(12.9^2/9)^2}{9-1}+\frac{(7.1^2/11)^2}{11-1}}\right] = 11, \quad t_{1-0.05}(11) = -t_{0.05}(11) = -1.796 \tag{6.4}$$

$$t = \frac{\bar{x}_1 - \bar{y}_2}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} = \frac{93 - 132}{\sqrt{\frac{12.9^2}{9} + \frac{7.1^2}{11}}} \approx -8.119 < t_{0.95} = -1.796 \Rightarrow \text{ Reject } H_0 \tag{6.5}$$

Exercise 6.2. Let Y be $b(192, p)$. We reject $H_0 : p = 0.75$ and accept $H_1 : p > 0.75$ if and only if $Y \geq 152$. Use the normal approximation to determine

1. $\alpha = P(Y \geq 152; p = 0.75)$.
2. $\beta = P(Y < 152)$ when $p = 0.80$.

Solution:

Proportion for one sample. $n = 192$

1. $\sum_{i=1}^{n} X_i = 152$, according to CLT and half-unit correction

$$z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{151.5 - 192(0.75)}{\sqrt{192(0.75)(1-0.75)}} \approx 1.25, z \overset{approx}{\sim} \mathcal{N}(0,1) \tag{6.6}$$

$$\alpha = P(Y \geq 152; p = 0.75) = P(Y > 151.5) = P(z > 1.25) = 0.1056 \tag{6.7}$$

2. $p = 0.8$ now, similarly,

$$z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{151.5 - 192(0.80)}{\sqrt{192(0.8)(1-0.8)}} \approx -0.38 \tag{6.8}$$

$$\beta = P(Y < 152) = P(Y < 151.5) = P(z < -0.38) = P(z > 0.38) = 0.3520 \tag{6.9}$$

Exercise 6.3. Let p equal the proportion of drivers who use a seat belt in a state that does not have a mandatory seat belt law. It was claimed that $p = 0.14$ An advertising campaign was conducted to increase this proportion. Two months after the campaign, $y = 104$ out of a random sample of $n = 590$ drivers were wearing their seat belts. Was the campaign successful?

1. Define the null and alternative hypotheses.
2. Define a critical region with an $\alpha = 0.01$ significance level. $z_{0.01} = 2.326$
3. What is your conclusion?

Solution:

1. $H_0 : p = 0.14$ against $H_1 : p > 0.14$
2. One sided proportion problem, $z_{0.01} = 2.326$.

$$C = \{z : z \geq 2.326\} \quad \text{where} \quad z = \frac{y/n - 0.14}{\sqrt{(0.14)(0.86)/n}} \tag{6.10}$$

3. For this problem, $y = 104, n = 590$, the value of test statistics is,

$$z = \frac{104/590 - 0.14}{\sqrt{(0.14)(0.86)/590}} = 2.539 > 2.326 \tag{6.11}$$

Hence, we reject H_0 and conclude that the advertising campaign indeed increases this proportion.

Exercise 6.4. For developing countries in Africa and the Americas, let p_1 and p_2 be the respective proportions of babies with a low birth weight (below 2500 grams). We shall test $H_0 : p_1 = p_2$ against the alternative hypothesis $H_1 : p_1 > p_2$

1. Define a critical region that has an $\alpha = 0.05$ significance level. $z_{0.05} = 1.645$
2. If respective random samples of sizes $n_1 = 900$ and $n_2 = 700$ yielded $y_1 = 135$ and $y_2 = 77$ babies with a low birth weight, what is your conclusion?
3. What would your decision be with a significance level of $\alpha = 0.01$? $z_{0.01} = 2.326$
4. What is the p-value of your test?

Solution:

Two samples proportion problem with $H_0 : p_1 = p_2$ against $H_1 : p_1 > p_2$.

1.
$$C = \{z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\,(1/n_1 + 1/n_2)}} \geq 1.645\} \tag{6.12}$$
where $\hat{p}_1 = y_1/n_1, \hat{p}_2 = y_2/n_2$, and $\hat{p} = \frac{y_1+y_2}{n_1+n_2}$.
2. Calculate the test statistic,

$$z = \frac{0.15 - 0.11}{\sqrt{(0.1325)(0.8675)(1/900 + 1/700)}} = 2.341 > 1.645 \tag{6.13}$$

Hence, we reject H_0 and conclude that the proportions of babies with a low birth weight in Africa is larger than that in Americas.

3. Since $z = 2.341 > 2.326 = z_{0.01}$, we reject H_0 and conclude that the proportions of babies with a low birth weight in Africa is larger than that in Americas.
4. The p-value is

$$P(z \geq 2.341) = 0.0096 \tag{6.14}$$

where z asymptotically follows $\mathcal{N}(0, 1)$.

Song, Wheyming Tina. 2005. "Relationships Among Some Univariate Distributions." *IIE Transactions* 37 (7): 651–56.