Resampling Techniques and their Application

-Class 3-

Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie

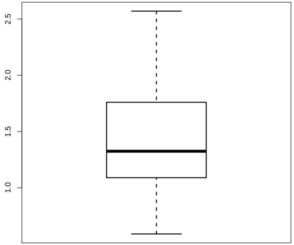
Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

Motivation and Examples-II

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of **n=36** bottles and obtains:

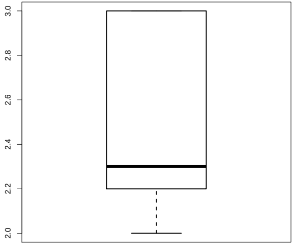



```
x=c(
0.59, 1.23, 1.00, 0.84, 0.88, 1.71,
1.81, 1.84, 2.03, 1.39, 1.30, 1.31,
1.96, 1.33, 2.57, 1.19, 1.01, 2.06,
1.32, 1.55, 1.28, 0.93, 1.63, 1.24,
1.83, 1.81, 0.94, 1.46, 1.25, 1.56,
0.61, 0.83, 1.17, 2.24, 1.68, 1.51)
```

- Data Analysis: Confidence interval and t-test

Motivation and Examples-III

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail.




```
x=c(
2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
```

- Data Analysis: Confidence interval and t-test

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$
- Test statistic: $T = \sqrt{36} * \frac{1.4136 - 1.5}{0.46} = -1.13$
- p-value: 2 times the area to the **left** of T under the t curve with 35 df. (Here, p=0.27)
- Quality of the estimator? Is the method valid? Is $p = 0.27$ a good estimate?

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$
- CI: $[2.21; 2.79]$
- Quality of the confidence interval?

Simulations

- With the help of simulations, we are able to
  - Study the behavior of point estimators
  - Visualize the distribution of the estimators
  - Assess parameters that are hard to estimate, e.g., variance of $\widehat{\sigma}^2$,
  - Assess (actual) type-I error rates under different scenarios
  - Assess the (actual) power of a test
  - Assess (actual) coverage probability of a confidence interval
  - ...

Sampling from Known Population-I

- Draw a **random sample** from a **population**
- Compute the statistic of interest

- Sample from Normal

```
set.seed(1)
x<-rnorm(10)
#10 draws
```

- Sample from Exponential

```
set.seed(1)
x<-rexp(10)
#10 draws
```

- Sample from Poisson

```
set.seed(1)
x<-rpois(10)
#10 draws
```

Sampling from Known Population-II

- Toss a coin (sides **H**ead and **T**ail). Each side is 50/50.
- ```
  set.seed(1)
  Coin <- c("H","T") #Population
  sample(Coin,1, replace=TRUE) #1 toss
  sample(Coin,2, replace=TRUE) #2 toss
  sample(Coin,3, replace=TRUE) #3 toss
  sample(Coin,10^6,replace=TRUE)#10^6 toss
  ```
- Set Head=1 and Tail=0 and repeat.
- Is the coin fair?

https://www.clipartkey.com/view/iRhJhxm_vector-illustration-of-decision-making-hand-flipping-probability/
  ```

Sampling from Known Population-III

- Roll a die. Each side is 1/6.
- ```
  set.seed(1)
  D <- c(1,2,3,4,5,6) #Population
  sample(D,1, replace=TRUE) #1 roll
  sample(D,2, replace=TRUE) #2 roll
  sample(D,3, replace=TRUE) #3 roll
  sample(D,10^6,TRUE)       #10^6 roll
  ```

- What is the expected value = mean of the population?
- What is the variance of the population?

https://www.aktivwelt.de/Freizeit-Fitness/Spiele/
  ```

Sampling from Known Population-IV

- Suppose we roll a die and let X denote the showing number. If the die is fair, then $\mu = E(X) = 3.5$. Suppose we roll the die $n = 10$ times (and obtain $X_1, \ldots, X_n$) and want to test the null hypothesis $H_0 : \mu = 3.5$. As test statistic we chose

$$T = \sqrt{n}\frac{\overline{X}_{\cdot} - 3.5}{\widehat{\sigma}}$$

- Compute the exact distribution of T and thus exact p-values
- How?
- Mathematically, or **by sampling**
  1. Generate a large number of samples from the population
  2. Compute values of the test statistics for each sample
  3. Compute the $(\alpha/2)$ and $(1 - \alpha/2)$ quantiles of the test statistics
  4. **We sample the distribution of the statistic**

Sampling from Known Population-V

```
D<-c(1,2,3,4,5,6) #population
nsim<- 10^7 #large number
n<-10
```

```
set.seed(1)
xR <- matrix(sample(D,n*nsim,TRUE),ncol=nsim)
mxR <- colMeans(xR)
sdxR <- sqrt((colSums(xR^2)-n*mxR^2)/(n-1))
TR <- sqrt(n)*(mxR-3.5)/sdxR
```

```
c1 <- quantile(TR,0.025) #2.5%quantile
c2 <- quantile(TR,0.975) #97.5%quantile
alphas <- seq(0,1,0.001)
distr <-quantile(TR,alphas)#distribution
hist(TR,freq=F)
```

Sampling from Known Population-VI

- Verify in a simulation study to test $H_0 : \mu = 3.5$ at 5% level

```
D<-c(1,2,3,4,5,6) #population
nsim<- 10^5 #large number
n<-10
set.seed(3)
x <- matrix(sample(D,n*nsim,TRUE),ncol=nsim)
mx <- colMeans(x)
sdx <- sqrt((colSums(x^2)-n*mx^2)/(n-1))
T <- sqrt(n)*(mx-3.5)/sdx
mean(abs(T)>= c2) #exact
mean(abs(T) >=qt(0.975,n-1)) #t test
```

- The true critical value is larger than the used *t*-quantile. Hence, the t-test over rejects.
  Note: The distribution is discrete.

Sampling from Known Populations-VII

- Suppose waiting times at a fast food restaurants follow Exponential distribution. We observe $n = 10$ customers (and obtain $X_1, \ldots, X_n$) and want to test the null hypothesis $H_0 : \mu = 1$. Compute the exact distribution (and thus p-values) of

$$T = \sqrt{n}\frac{\overline{X}_{\cdot} - 1}{\widehat{\sigma}}$$

```
nsim <- 10^7 #large number
n<-10; set.seed(1)
xR <- matrix(rexp(n*nsim),ncol=nsim);mxR <- colMeans(xR)
sdxR <-sqrt((colSums(xR^2)-n*mxR^2)/(n-1));TR <- sqrt(n)*(mxR-1)/sdxR
c1 <- quantile(TR,0.025) #2.5%quantile
c2 <- quantile(TR,0.975) #97.5%quantile
alphas <- seq(0,1,0.001)
distr <-quantile(TR,alphas)#distribution
hist(TR,freq=F)
```

Sampling from Known Populations-VIII

- Suppose waiting times at a fast food restaurants follow Exponential distribution. Suppose we observe $n = 10$ customers (and obtain $X_1, \ldots, X_n$) and want to test the null hypothesis $H_0 : \mu = 1$. Compute the exact distribution (and thus p-values) of

  $$T = \sqrt{n}\frac{\overline{X}_{\cdot} - 1}{\widehat{\sigma}}$$

  The exact 2.5% and 97.5% critical values are $c1$ and $c2$ (see above). Check in a type-I error simulation.

```
nsim <- 10^7 #large number
n<-10; set.seed(3)
x <- matrix(rexp(n*nsim),ncol=nsim);mx <- colMeans(x)
sdx <-sqrt((colSums(x^2)-n*mx^2)/(n-1))
T <- sqrt(n)*(mx-1)/sdx
mean(T<c1 | T >c2) #exact
mean(abs(T)>=qt(0.975,n-1)) #ttest
```

Sampling from Known Population-IX

- Revise the cork diameter example. Assume **sample=Population**.
`X=c(0.59, 1.23, 1.00, 0.84, 0.88, 1.71,1.81, 1.84, 2.03, 1.39, 1.30, 1.31, 1.96, 1.33, 2.57, 1.19, 1.01, 2.06,1.32, 1.55, 1.28, 0.93, 1.63, 1.24, 1.83, 1.81, 0.94, 1.46, 1.25, 1.56, 0.61, 0.83, 1.17, 2.24, 1.68, 1.51)`
- If this sample is the population and we draw X^* with replacement from X. What is $E(X^*)$?

$$\begin{aligned} P(X^* &= X_k) = \frac{1}{n} = \frac{1}{36} \\ E(X^*) &= \sum_{k=1}^{n} X_k * P(X^* = X_k) = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}. \end{aligned}$$

- On average, we observe $\overline{X}$.

Sampling from Known Population-X

- Revise the cork diameter example. Assume **sample=Population.**
- Whichever parameter we would like to estimate. How can we easily do it?
- We sample from X and find the information
  - Suppose we draw n observations from X and obtain $X_1^*, \ldots, X_n^*$ with $E(X_k^*) = \mu = \overline{X}_.$
    - $\overline{X}_.^* = \frac{1}{n}\sum_{k=1}^n X_k^*$
    - $\widehat{\sigma}_*^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k^* - \overline{X}_.^*)^2$
  - Compute the distribution of

$$T^* = \sqrt{n}\frac{\overline{X}_.^* - \overline{X}_.}{\widehat{\sigma}_*}$$

Sampling from Known Population-XI

```
nsim<- 10^7 #large number
n<-10
set.seed(1)
xR <- matrix(sample(X,n*nsim,TRUE),ncol=nsim)
mxR <- colMeans(xR)
sdxR <-sqrt((colSums(xR^2)-n*mxR^2)/(n-1))
TR <- sqrt(n)*(mxR-mean(X))/sdxR
c1 <- quantile(TR,0.025) #2.5%quantile
c2 <- quantile(TR,0.975) #97.5%quantile
alphas <- seq(0,1,0.001)
distr <-quantile(TR,alphas)#distribution
hist(T)
```

Sampling from Sample (Nonparametric Bootstrap)

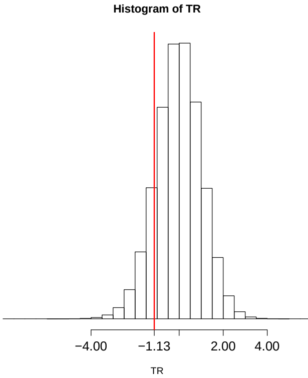
- So far so good, but...
- **The population is unknown**
- How does the above help us?
- **We fix the sample and treat it like a population**
- Logical Flow
  1. Fix the data (size n)
  2. Draw n variables with replacement from the data
  3. Compute the statistic
  4. Repeated the last two steps a large number of times and compute statistic of interest

Sampling from Sample (Nonparametric Bootstrap)-II

- Revise the cork diameter example. Now we **treat it as a sample**
- We computed the test statistic as $T = -1.13$.
- **Given the sample**, compute the resampling distribution of T

```
nsim<- 10^6 #large number
n<-36
set.seed(1)
Torig=-1.13 #test statistic
```

```
xR <- matrix(sample(X,n*nsim,TRUE),ncol=nsim)
mxR <- colMeans(xR)
sdxR <-sqrt((colSums(xR^2)-n*mxR^2)/(n-1))
TR <- sqrt(n)*(mxR-mean(X))/sdxR
```

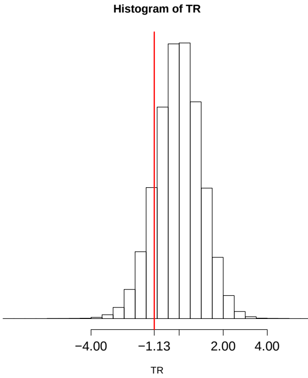


Sampling from Sample (Nonparametric Bootstrap)-III

- Revise the cork diameter example. Now we **treat it as a sample**
- We computed the test statistic as $T = -1.13$.
- **Given the sample**, compute the critical values (5%) and p-value

```
nsim<- 10^6 #large number
n<-36;set.seed(1)
Torig=-1.13 #test statistic
```

```
TR <- sqrt(n)*(mxR-mean(X))/sdxR
c1 <- quantile(TR,0.025);c2 <-quantile(TR,0.975)
#pvalue: %of TR more extreme than Torig
#in either direction
pvalue<-2*min(mean(TR<=Torig), mean(TR>=Torig))
```

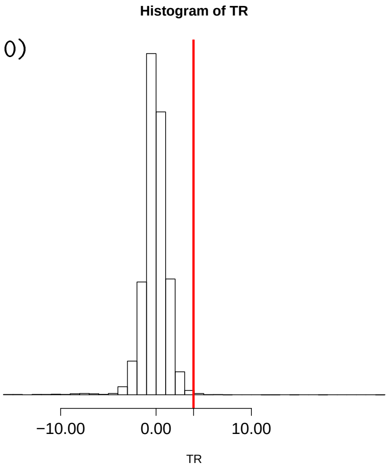


- ttest: $pvt < -2 * min(pt(Torig, n-1), 1 - pt(Torig, n-1))$

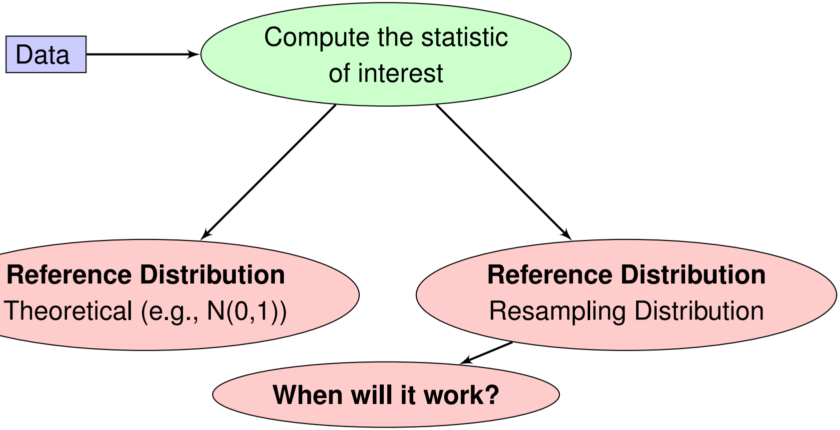
(Nonparametric Bootstrap)-IV

- Revise the reaction time data
- We aim to test $H_0 : \mu = 2$. Compute a resampling test and t-test.

- ```
  x<-c(2.4, 3.0, 3.0, 2.2, 2.2, 2.2, 2.2, 2.8, 2.0, 3.0)
  n<-10;set.seed(1)
  mx <- mean(x)
  sdx <-sd(x)
  T<-sqrt(n)*(mx-2)/sdx  #original
  #----------------------------------------
  nsim<- 10^6 #large number
  xR <- matrix(sample(x,n*nsim,TRUE),ncol=nsim)
  mxR <- colMeans(xR)
  sdxR <-sqrt((colSums(xR^2)-n*mxR^2)/(n-1))
  TR <- sqrt(n)*(mxR-mx)/sdxR
  pvalue<-2*min(mean(TR<=T), mean(TR>=T))
  pvt <-2*min(pt(T,n-1),1-pt(T,n-1))#ttest
  ```


  ```

The Resampling Work Flow



Project: t-test versus Resampling

- We draw the random sample $X : X_1, \ldots, X_n$ with $E(X) = \mu$ and variance σ^2 from a population and we aim to test the null hypothesis $H_0 : \mu = 0$.
- We use the test statistic

$$T = \sqrt{n}\frac{\overline{X}.}{\widehat{\sigma}}$$

- Should we use a t-test or Resampling test? (p-values from $t(n-1)$ or resampling)
- Perform type-I error simulations for normal and exponential distributions with varying $n \in \{10, 15, 20\}$ at $\alpha = 5\%$ level of significance ($n_{sim} = 10,000$). Use set.seed(1)
- Quality criteria: Use the **precision interval** for the estimated type-I error rate, $\widehat{\alpha}$,

$$PI(\alpha, n_{sim}) = \left[\alpha - \frac{1.96}{n_{sim}}\sqrt{\alpha(1-\alpha)}; \alpha + \frac{1.96}{n_{sim}}\sqrt{\alpha(1-\alpha)}\right]$$

- We call the procedure *accurate*, if $\widehat{\alpha} \subseteq PI_{\alpha, n_{sim}}$. If $\widehat{\alpha}$ exceeds the upper bound, the test is called *liberal*, otherwise *conservative*. Which method do you recommend?