1. **Problem**
I have a sample $X_1, \ldots, X_{60}$. I generate one naive bootstrap sample $X_1^*, \ldots, X_{60}^*$.
Let N be the number of times the first observation will be copied in the bootstrap sample.
Find the probability $\mathbb{P}(N = 2)$.

SRS with replacement

$X_1 \rightrightarrows h \quad h^*$

$\frac{1}{60}$	$\frac{1}{60}$	$\frac{59}{60}$		$\frac{59}{60}$
1	2	3	....	58

$X_1 \nearrow 1/60$

$X_1 \searrow 59/60$

$$P(N=2) = \frac{1}{60} \cdot \frac{1}{60} \cdot \left(\frac{59}{60}\right)^{58} \cdot C_{60}^{2}$$

$$N \sim \text{Binom}(n, p) = \text{Bin}\left(60, \frac{1}{60}\right)$$

$$P(N=k) = C_n^k \cdot p^k \cdot (1-p)^{n-k}$$

$$P(N \le 1) = P(N=0) + P(N=1)$$

2. **Problem**
I have a sample $X_1, \ldots, X_{60}$. I generate one naive bootstrap sample $X_1^*, \ldots, X_{60}^*$.
Let L be the number of initial observations missing in the bootstrap sample.
Find the expected value $\mathbb{E}(L)$.

$$l_i = \begin{cases} 1, & \text{missing} \\ 0, & \text{not missing} \end{cases} \qquad \begin{matrix} 59/60 \\ 1/60 \end{matrix}$$

$$\sum l_i = L$$

$$E(l_i) = p = \frac{59}{60} \qquad Var(l_i) = pq = \frac{59}{60} \cdot \frac{1}{60}$$

$$E(L) = \underset{X_1}{1 \cdot \left(\frac{59}{60}\right)^{60}} + \underset{X_2}{1 \cdot \left(\frac{59}{60}\right)^{60}} + \ldots + \underset{X_{60}}{1 \cdot \left(\frac{59}{60}\right)^{60}}$$

$$= 60 \cdot \left(\frac{59}{60}\right)^{60}$$

3. **Problem**
We have data of an AB experiment: $\bar{X}_a = 5.1$, $\bar{X}_b = 5.6$, $n_a = 18$, $n_b = 12$, $\sum(X_i^a - \bar{X}_a)^2 = 840$, $\sum(X_i^b - \bar{X}_b)^2 = 820$.
Calculate the estimate of variance of $\bar{X}_a - \bar{X}_b$ for the Welch test.

t-test vs Welch Test

$H_0: \mu_a = \mu_b$ $\quad\quad$ $H_0: \mu_a = \mu_b$

Ass.: $\sigma_a = \sigma_b = \sigma_p$ $\quad\quad$ Ass.: $\sigma_a \neq \sigma_b$

$\hat{v} \leq n-2$

$$t = \frac{\bar{X}_a - \bar{X}_b - 0}{se(\bar{X}_a - \bar{X}_b)} \sim t_{n-2} \quad\quad t = \frac{\bar{X}_a - \bar{X}_b - 0}{se(\bar{X}_a - \bar{X}_b)} \sim t_{\hat{v}}$$

$$t = \frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

$$se(\bar{X}_a - \bar{X}_b) = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \quad\quad se(\bar{X}_a - \bar{X}_b) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$S_p^2 = \frac{\overbrace{(n_1 - 1) S_1^2}^{TSS_1} + \overbrace{(n_2 - 1) \cdot S_2^2}^{TSS_2}}{n_1 + n_2 - 2}$$

$$\hat{\sigma}_1^2 = \frac{1}{n} \cdot \sum (X_i - \bar{X})^2$$

$$S_1^2 = \frac{1}{n-1} \cdot \sum (X_i - \bar{X}_1)^2 = \frac{1}{n_1 - 1} TSS_1$$

$$S_2^2 = \frac{TSS_2}{n_2 - 1}$$

$$\hat{Var}(\bar{X}_a - \bar{X}_b) = \sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}} = \sqrt{\frac{840/17}{18} + \frac{820/11}{12}}$$

3. **Problem**
We have data of an AB experiment: $\bar{X}_a = 5.1$, $\bar{X}_b = 5.6$, $n_a = 18$, $n_b = 12$, $\sum(X_i^a - \bar{X}_a)^2 = 840$, $\sum(X_i^b - \bar{X}_b)^2 = 820$.
Calculate the estimate of variance of $\bar{X}_a - \bar{X}_b$ for the Welch test.

4. **Problem**

Variables X_1, X_2, ..., X_{20} are independent and normally distributed $\mathcal{N}(5;3)$.

Calculate the variance

$$\operatorname{Var}\left(\sum_{i=1}^{20}(X_i-\bar{X})^2\right).$$

$X_i \sim N(5,3)$ $\quad 3 = \sigma^2$ $\quad \frac{\sum X_i}{n}$

$Cov(X_i - \bar{X}, X_j - \bar{X}) \neq 0$

$Var(TSS) - ?$

$\sum \frac{(X_i - \bar{X})^2}{\sigma^2} = \sum \xi^2$

$\chi^2_k = \sum_{i=1}^{k} \xi_i^2 \quad , \quad \xi_i \sim N(0,1)$

$s^2 = \frac{1}{n-1} \cdot TSS$

$\hookrightarrow \quad n \cdot R^2 \sim \chi^2$

$\hookrightarrow \quad F = \frac{Var(\cdot)}{Var(\cdot)} \sim \frac{\chi^2_p}{\chi^2_q} = F_{p,q}$

$\frac{(n-1)\, s^2}{\sigma^2} \sim \chi^2_{n-1}$

$Var\left(\frac{n-1}{\sigma^2} s^2\right) = Var(\chi^2_{n-1})$

$\frac{(\cancel{n-1})^2}{\sigma^4} \cdot Var\left(\frac{TSS}{\cancel{n-1}}\right) = Var(\chi^2_{n-1})$

$E(\xi) = 0$

$E(\xi^2) = Var(\xi)$

$Var(\chi^2_1) = Var(\xi^2) =$

$E(\xi^4) - E^2(\xi^2) = \mu_4 - \mu_2^2 =$

$3 - 1 = 2$

$Var(\chi_k) = Var\left(\sum_{i=1}^{k} \xi_i^2\right) = \sum_{i=1}^{k} Var(\xi_i^2) =$

$= k \cdot 2$

$\frac{Var(TSS)}{\sigma^4} = (n-1)\, 2$

$Var(TSS) = 2 \cdot (n-1) \cdot \sigma^4 =$

$= 2 \cdot 19 \cdot (\sqrt{3})^4$

5. **Problem**

I have results of two runners A and B: 2 results for A and 3 results for B. The running time for both runners are continuously distributed and their distribution are equal.

What is the probability that the maximal rank of running times of the runner A will be equal to 4?

$X^A \sim N(\mu_a, \sigma_a^2)$

① $\begin{cases} A: & 2 \quad 4 \\ B: & 3 \quad 5 \quad 6 \end{cases}$

$\text{rank}(X_3^B) = 5$

A ? ? A B

1 2 3 4 5

C_1^3 : ABB BAB BBA

$$P(\max(\text{rank } A) = 4) = \frac{C_1^3}{C_2^5}$$

$$p(\max(\text{rank } A) \le 4)$$

rank = 1 ? ? A B B

A

rank = 2 A A B B B

$p(\text{rank} = 1) = 0$

6. **Problem**

I have results of two runners A and B: 2 results for A and 3 results for B. The running times for both runners are continuously distributed.

Consider the Mann-Whitney test statistic U_a that positively depends on the rank sum for the runner A.

Find the expected value $\mathbb{E}(A)$ under the null-hypothesis of equal running time distributions.

T/W $H_0: \mu_a = \mu_b$

MW $H_0: E(\text{rank}_a) = E(\text{rank}_b)$

$\hookrightarrow H_0: f_a(x) = f_b(x+c),\ c=0$

$\hookrightarrow H_0: P(X \le Y) = 0{,}5$

$\hookrightarrow H_0: \mu_a = \mu_b$

A : 2 4

B : 3 5 6

	2	3	4	5	6	
A	0		1			$\Sigma \Rightarrow U_a = 1$
B		1		2	2	$\Sigma \Rightarrow U_b = 5$

$$U = \min\{U_a, U_b\} = 1$$

$$U \overset{d}{\sim} N\left(\frac{n_1 n_2}{2}; \frac{n_1 n_2 (n_1 + n_2 - 1)}{12}\right)$$

$U_{obs} = 1$ vs U_{crit}

$$\frac{U_{obs} - E(U)}{sd(U)} \sim N(0,1)$$

$$\frac{1 - \frac{2 \cdot 3}{2}}{\sqrt{\frac{2 \cdot 3 \cdot 4}{12}}} = \frac{-2}{\sqrt{2}} \approx -1{,}4$$

$|-1{,}4| < 1{,}96$

$\Rightarrow H_0$ is not rej.

7. **Problem**
Consider three variables: target variable y_i, predictor x_i and indicator of treatment $z_i \in \{0, 1\}$.

The treatment z_i was designed to be independent of x_i, but in fact $x_i = f_i \cdot (1 + 0.08z_i)$.

We suppose that z_i are Bernoulli with $p = 0.4$, $f_i \sim \mathcal{N}(900; 9)$ and they are independent.

Find the probability limit

$$\text{plim} \frac{\sum(x_i - \bar{x})(z_i - \bar{z})}{n - 1}.$$

$\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

$$\text{plim}_{n\to\infty} \frac{\bar{X}}{n} = E(X) = \mu$$

$$\text{plim}_{n\to\infty} \sum (X_i - \bar{X})^2 = \infty$$

$$\text{plim}_{n\to\infty} \frac{\sum (X_i - \bar{X})^2}{n} = \text{plim}_{n\to\infty} \frac{\sum (X_i - \bar{X})}{n-1} = \mathrm{Var}(X)$$

7. **Problem**
Consider three variables: target variable y_i, predictor x_i and indicator of treatment $z_i \in \{0, 1\}$.

The treatment z_i was designed to be independent of x_i, but in fact $x_i = f_i \cdot (1 + 0.08z_i)$.

We suppose that z_i are Bernoulli with $p = 0.4$, $f_i \sim \mathcal{N}(900; 9)$ and they are independent.

Find the probability limit

$$\text{plim} \frac{\sum(x_i - \bar{x})(z_i - \bar{z})}{n - 1}.$$

$$\text{plim}_{n\to\infty} \widehat{\mathrm{Cov}}(X, Z) = \mathrm{Cov}(X_i, Z_i) = \mathrm{Cov}(f_i \cdot (1 + 0{,}08 z_i), z_i)$$

$$= \underbrace{\mathrm{Cov}(f_i, z_i)}_{=0} + 0{,}08 \cdot \mathrm{Cov}(f_i z_i, z_i) =$$

$z_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$

$z_i^2 = \begin{cases} 1^2 \\ 0^2 \end{cases}$

$$= 0{,}08 \cdot \left(E(f_i z_i^2) - E(f_i z_i) \cdot E(z_i) \right) =$$

$$= 0{,}08 \cdot E(f_i z_i) \left(1 - E(z_i)\right) =$$

$$= 0{,}08 \cdot 900 \cdot 0{,}4 \cdot (1 - 0{,}4)$$

8. **Problem**

Consider three variables: target variable y_i, predictior x_i and indicator of treatment $z_i \in \{0,1\}$. The treatment z_i was assigned independently of x_i, total $n = 100$.

The sample covariance matrix C is provided. The order of variables is y, x and z. For example the sample covariance of x and z is 0.26.

$$C = \begin{array}{c} \\ y \\ x \\ z \end{array} \begin{array}{c} \begin{array}{ccc} y & x & z \end{array} \\ \begin{pmatrix} 7.09 & 2.64 & 0.75 \\ 2.64 & 1.17 & 0.26 \\ 0.75 & 0.26 & 0.25 \end{pmatrix} \end{array}$$

Consider CUPED with first regression given by $\hat{y}_i = \hat{\alpha}_1 + \hat{\alpha}_2 x_i$ with residuals $\underset{y_{cuped}}{r_i} = y_i - \hat{y}_i$.

What is the sample covariance of r and z?

Cuped: 1) $Y_{cuped} = Y - \theta X$

$$\hat{\theta} = \frac{\hat{Cov}(X,Y)}{\hat{Var}(X)} \Leftrightarrow \hat{\beta} \text{ in } Y|X$$

2) $Var(\bar{Y}_{cuped}) = (1 - \rho^2_{X,Y}) Var(\bar{Y}) \leq Var(\bar{Y})$

$$\rho^2_{X,Y} = R^2_{Y|X}$$

$$\hat{Cov}(r,z) = \hat{Cov}(y - \hat{y}, z) =$$

$$= \hat{Cov}(Y - \hat{\theta} X, z) =$$

$$\hat{Cov}(Y,z) - \hat{\theta}\,\hat{Cov}(X,z) =$$

$$\hat{Cov}(Y,z) - \frac{\hat{Cov}(X,Y)}{\hat{Var}(X)} \hat{Cov}(X,z) =$$

$$0{,}75 - \frac{2{,}64}{1{,}17} \cdot 0{,}26$$