Estimating a relative change using a log-transformation of the outcome

Brice Ozenne

March 8, 2021

1 Interpretation of the regression coefficient after log-transformation

Let's denote by Y the outcome and by G a binary group variable. We are interested in the relative change in Y between the groups. We decide to model the group effect on the log scale:

$$\log(Y) = Z = \alpha + \beta G + \varepsilon \text{ where } \mathbb{E}\left[\varepsilon\right] = 0 \text{ and } \mathbb{E}\left[\varepsilon\right] = \sigma^2$$

We claim that:

$$\frac{\mathbb{E}\left[Y|G=1\right] - \mathbb{E}\left[Y|G=0\right]}{\mathbb{E}\left[Y|G=0\right]} = e^{\beta} - 1$$

1.1 Proof: re-writting the model as a multiplicative model

We can re-write the model as:

$$Y = e^{\alpha+\beta G} e^{\varepsilon} \text{ where}$$

So for $g \in \{1, 2\}$:

$$\mathbb{E}\left[Y|G=g\right] = e^{\alpha+\beta g}\mathbb{E}\left[e^{\varepsilon}\right]$$

Then:

$$\begin{aligned}
\frac{\mathbb{E}\left[Y|G=1\right] - \mathbb{E}\left[Y|G=0\right]}{\mathbb{E}\left[Y|G=0\right]} &= \frac{e^{\alpha+\beta}\mathbb{E}\left[e^{\varepsilon}\right] - e^{\alpha}\mathbb{E}\left[e^{\varepsilon}\right]}{e^{\alpha}\mathbb{E}\left[e^{\varepsilon}\right]} \\
&= \frac{e^{\alpha+\beta} - e^{\alpha}}{e^{\alpha}} = e^{\beta} - 1
\end{aligned}$$

1.2 Proof: using a Taylor expansion

Using a second order Taylor expansion of $\exp(Z)$ around $\mu(G) = \alpha + \beta G$ and assuming that the first moments of Z are finite and the remaining moments are neglectable regarding the factorial of the moment order (i.e. $\forall i \geq 1$, $\frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right] < +\infty$ and $\sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right] < +\infty$), we get:

$$\begin{aligned}
Y = e^Z &= e^\mu + \sum_{i=1}^{\infty} \frac{1}{i!}(Z-\mu)^i \frac{\partial^i e^\mu}{(\partial\mu)^i} \\
&= e^{\alpha+\beta G} + \sum_{i=1}^{\infty} \frac{1}{i!}(Z-\alpha-\beta G)^i e^{\alpha+\beta G} \\
\mathbb{E}\left[Y|G=g\right] &= e^{\alpha+\beta G} + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[(Z-\alpha-\beta g)^i\right] e^{\alpha+\beta G} \\
&= e^{\alpha+\beta G}\left(1 + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right]\right)
\end{aligned}$$

where we used that the distribution of ε is independent of g. We can now express our parameter of interest:

$$\begin{aligned}
\Delta_G &= \frac{\mathbb{E}\left[Y|G=1\right] - \mathbb{E}\left[Y|G=0\right]}{\mathbb{E}\left[Y|G=0\right]} = \frac{\mathbb{E}\left[Y|G=1\right]}{\mathbb{E}\left[Y|G=0\right]} - 1 \\
&= \frac{e^{\alpha+\beta}\left(1 + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right]\right)}{e^{\alpha}\left(1 + \sum_{i=1}^{\infty} \frac{1}{i!}\mathbb{E}\left[\varepsilon^i\right]\right)} - 1 \\
&= e^\beta - 1
\end{aligned}$$

2 Power calculation: comparison between two groups

Consider two groups $G = 0$ and $G = 1$ for which we want to compare the percentage difference in outcome Y. We are willing to assume that on the log-scale Y is normally distributed. Our parameter of interest is:

$$\frac{\mathbb{E}\left[Y|G=1\right] - \mathbb{E}\left[Y|G=0\right]}{\mathbb{E}\left[Y|G=0\right]} = \gamma$$

We further fix $\alpha = \mathbb{E}\left[Y|G=0\right]$ and $\sigma^2 = \mathbb{V}ar\left[Y|G=0\right]$ and we assume that on the log-scale:

$$\mathbb{V}ar\left[\log(Y)|G=1\right] = \mathbb{V}ar\left[\log(Y)|G=0\right] = s^2$$

To evaluate the power for a given $(\alpha, \sigma^2, \gamma)$, we need to identify the joint distribution:

$$\begin{bmatrix} Z_0 = \log(Y)|G=0 \\ Z_1 = \log(Y)|G=1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_0 \\ m_1 \end{bmatrix}, \begin{bmatrix} s^2 & \rho s^2 \\ \rho s^2 & s^2 \end{bmatrix}\right)$$

The standardized effect size is then: $\frac{m_1 - m_0}{s\sqrt{2(1-\rho)}}$.

Note: in the case of two independent samples $\rho = 0$

2.1 Method 1: Taylor expansion

We will use the fact that Z_0, Z_1 are jointly normally distributed to identify m_0, m_1, s^2, ρ. First we start by identifying m_0 and s^2 based on α and σ^2 (reference group). A Taylor expansion gives (see appendix B.2):

$$\begin{aligned} \alpha &\approx \exp(m_0)\left(1 + \frac{s^2}{2} + \frac{s^4}{8} + \frac{s^6}{48}\right) \\ \sigma^2 &\approx \exp(2m_0)\left(s^2 + \frac{3}{2}s^4 + \frac{7}{6}s^6 + \frac{11}{24}s^8 + \frac{21}{320}s^{10}\right) \end{aligned}$$

So:

$$\frac{\alpha^2}{\sigma^2} - \frac{\left(1 + \frac{s^2}{2} + \frac{s^4}{8} + \frac{s^6}{48}\right)^2}{s^2 + \frac{3}{2}s^4 + \frac{7}{6}s^6 + \frac{11}{24}s^8 + \frac{21}{320}s^{10}} \approx 0$$

We get s^2 by solving this equation using that $s^2 \in [0; \sigma^2]$ (upper bound follow from Jensen's inequality applied to $(X-\mu)^2$, log being concave). We can then deduce m_0:

$$m_0 \approx \log\left(\frac{\alpha}{1 + \frac{s^2}{2} + \frac{s^4}{8} + \frac{s^6}{48}}\right) = \log(\alpha) - \log\left(1 + \frac{s^2}{2} + \frac{s^4}{8} + \frac{s^6}{48}\right)$$

Then we can identify m_1 using once more a Taylor expansion:

$$\alpha(1+\gamma) \approx \exp(m_1)\left(1+\frac{s^2}{2}+\frac{s^4}{8}+\frac{s^6}{48}\right)$$
$$m_1 \approx \log\left(\frac{\alpha(1+\gamma)}{1+\frac{s^2}{2}+\frac{s^4}{8}+\frac{s^6}{48}}\right) = m_0 + \log(1+\gamma)$$

Now

2.2 Method 2: Log-normal distribution

We will use the fact that Z_0 follows a log-normal distribution, meaning that:

$$\alpha = \exp(m_0 + \frac{1}{2}s^2)$$
$$\sigma^2 = \exp(2 * m_0 + s^2) * (\exp(s^2) - 1)$$

So

$$s^2 = \log\left(1+\frac{\sigma^2}{\alpha^2}\right)$$
$$m_0 = \log(\alpha) - \frac{s^2}{2}$$

Then we can identify m_1 using that Z_1 follows a log-normal distribution, i.e.:

$$\alpha(1+\gamma) = \exp(m_1 + \frac{1}{2}s^2)$$
$$m_1 = m_0 + \log(1+\gamma)$$

2.3 Illustration 1: two sample t-test

Illustration: We consider two groups having a 10% difference in their baseline value ($\alpha = 1.15$) and a variance of $\sigma^2 = 0.15$. What are the parameters of the corresponding normal distribution on the log-scale and the standardized effect size?

```
alpha <- 1.15
sigma2 <- 0.15
gamma <- 0.1
```

Taylor expansion: we first identify s^2, m_0, and m_1:

```
s2.taylor <- uniroot(function(x){
    alpha^2/sigma2 - (1+x/2+x^2/8+x^3/48)^2/(x+(3/2)*x^2+(7/6)*x^3+(11/24)
   *x^4+(21/320)*x^5)},
    interval = c(1e-12,sigma2))$root
m0.taylor <- log(alpha/(1+s2.taylor/2+s2.taylor^2/8+s2.taylor^3/48))
m1.taylor <- m0.taylor + log(1+gamma)
```

lognormal distribution: we first identify s^2, m_0, and m_1:

```
s2.logdist <- log(1+sigma2/alpha^2)
m0.logdist <- log(alpha) - s2.logdist/2
m1.logdist <- m0.logdist + log(1+gamma)
```

We can compare the moments of am exp-transformed normal distribution based on these values to the input:

```
x <- exp(rnorm(1e5, mean = m0.taylor, sd = sqrt(s2.taylor)))
y <- exp(rnorm(1e5, mean = m1.taylor, sd = sqrt(s2.taylor)))
yx.x <- mean(y)/mean(x)-1
X <- exp(rnorm(1e5, mean = m0.logdist, sd = sqrt(s2.logdist)))
Y <- exp(rnorm(1e5, mean = m1.logdist, sd = sqrt(s2.logdist)))
YX.X <- mean(Y)/mean(X)-1

rbind(data.frame(method = "taylor",
   m0=m0.taylor, m1=m1.taylor, s2=s2.taylor),
      data.frame(method = "logdist",
   m0=m0.logdist, m1=m1.logdist, s2=s2.logdist)
      )
rbind(data.frame(method = "true",
   alpha=alpha, gamma=gamma, sigma2=sigma2),
      data.frame(method = "error.taylor",
   alpha=mean(x)-alpha, gamma=yx.x-gamma, sigma2=var(x)-sigma2),
      data.frame(method = "error.logdist",
   alpha=mean(X)-alpha, gamma=YX.X-gamma, sigma2=var(X)-sigma2)
      )
```

```
   method         m0        m1        s2
1  taylor 0.08603197 0.1813421 0.1074606
2 logdist 0.08604307 0.1813532 0.1074378
        method         alpha         gamma        sigma2
1         true  1.1500000000  0.1000000000  0.1500000000
2  error.taylor  0.0012850559 -0.0010820104 -0.0002242144
3 error.logdist -0.0005174973 -0.0009134562 -0.0012306318
```

Similar performance. Maybe a bit better for log-dist.

2.4 Illustration 2: paired t-test

Illustration: We consider one group having a 10% difference between its baseline value ($\alpha = 1.15$) and its follow-up value. We assume a variance of $\sigma^2 = 0.15$ for the baseline value and a correlation of $\rho = 0.5$ between the baseline and follow-up value. What are the parameters of the corresponding normal distribution on the log-scale and the standardized effect size?

```
alpha <- 1.15
sigma2 <- 0.15
gamma <- 0.1
rho <- 0.5
```

We previously obtained the values fors^2. We can now search for the right correlation value on the log-scale

```
rho.taylor <- uniroot(function(x){
    rho - (x+1.5*x^2*s2.taylor+(1/12)*s2.taylor^2*(2*x^3+3*x))/(1+(3/2)*s2
    .taylor+(7/6)*s2.taylor^2+(11/24)*s2.taylor^3+(21/320)*s2.taylor^4)
},interval = c(0,0.9999))$root
```

```
library(mvtnorm)
Sigma <- diag(s2.taylor*(1 - rho.taylor),2,2)+s2.taylor*rho.taylor
z <- exp(rmvnorm(1e5, mean = c(m0.taylor, m1.taylor), sigma = Sigma))
c("true" = rho,
  "error.taylor" = rho-cor(z[,1],z[,2]))
```

```
      true error.taylor
0.50000000  -0.02621529
```

2.5 Application: two independent groups

We consider two groups having a 10% difference in their baseline value ($\alpha = 1.15$) and a variance of $\sigma^2 = 0.15$. What are the parameters of the corresponding normal distribution on the log-scale and the standardized effect size?

```
alpha <- 1.15
sigma2 <- 0.15
gamma <- 0.1
```

Solve the equations:

```
        a0         s0         a1         s1
0.08802784 0.10608948 0.19175319 0.08851048
```

We can check that `uniroot` converged correctly:

```
c(exp(a0)*(1+s0/2) - alpha,
  exp(2*a0)*(s0+s0^2*7/4) - sigma2,
  exp(a1)*(1+s1/2) - alpha*(1+gamma),
  exp(2*a1)*(s1+s1^2*7/4) - sigma2)
```

```
[1] -5.563198e-05  0.000000e+00 -1.895835e-05  0.000000e+00
```

and the variables have the appropriate distribution:

```
Z0 <- exp(rnorm(1e4, mean=a0, sd = sqrt(s0)))
Z1 <- exp(rnorm(1e4, mean=a1, sd = sqrt(s1)))
c(alpha = mean(Z0),
  gamma = (mean(Z1)-mean(Z0))/mean(Z0),
  sigma2 = var(Z0),
  sigma2 = var(Z1))
```

```
    alpha     gamma    sigma2    sigma2
1.1435272 0.1090391 0.1473705 0.1507638
```

For a power calculation we would use:

```
pwr.t.test(d = (a1-a0)/sqrt(s0/2+s1/2), sig.level = 0.05, power = 0.8)
## dvmisc::power_2t_unequal(n = 143, d = a1-a0, sigsq1 = s0, sigsq2 = s1,
    alpha = 0.05)
```

```
     Two-sample t test power calculation

              n = 142.9312
              d = 0.3325282
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

A Moments of the normal distribution

Denote X and Y two normally distributed variables, with mean μ_X,μ_Y and variance σ_X^2,σ_Y^2. Then:

- $\mathbb{E}\left[X^2\right] = \sigma_X^2 + \mu_X^2$
- $\mathbb{E}\left[X^3\right] = 3\mu_X\sigma_X^2 + \mu_X^3$
- $\mathbb{E}\left[X^4\right] = 3\left(\sigma_X^2\right)^2 + 6\sigma_X^2\mu_X^2 + \mu_X^4$
- $\mathbb{E}\left[X^5\right] = 15\left(\sigma_X^2\right)^2\mu + 10\sigma_X^2\mu^3 + \mu^5$
- $\mathbb{E}\left[(X-\mu_X)^6\right] = 15\left(\sigma_X^2\right)^3$
- $\mathbb{E}\left[(X-\mu_X)^8\right] = 105\left(\sigma_X^2\right)^4$

- $\mathbb{C}ov\left[X^2, X\right] = 2\mu_X\sigma_X^2$
- $\mathbb{C}ov\left[X^2, Y\right] = 2\mu_X\rho\sigma_X\sigma_Y$
- $\mathbb{E}\left[X^2 * Y^2\right] = (\sigma_X^2 + \mu_X^2)(\sigma_Y^2 + \mu_Y^2) + 2\rho^2\sigma_X^2\sigma_Y^2 + 4\rho\sigma_Y\sigma_X\mu_X\mu_Y$
- $\mathbb{C}ov\left[(X-\mu_X)^2, (Y-\mu_Y)^2\right] = 2\rho^2\sigma_X^2\sigma_Y^2$
- $\mathbb{C}ov\left[(X-\mu_X), (Y-\mu_Y)^3\right] = 3\rho\sigma_X\sigma_Y^3$
- $\mathbb{C}ov\left[(X-\mu_X)^3, (Y-\mu_Y)^3\right] = (6\rho^3 + 9\rho)\sigma_X^3\sigma_Y^3$

B Moments after transformation

B.1 Recall: Taylor expansion for normally distributed variables

Taylor expansion for a smooth function f around the mean value $\mu_Y = \mathbb{E}[Y]$:

$$f(Y) = f(\mu_Y) + f'(\mu_Y)(Y - \mu_Y) + \frac{1}{2}f''(\mu_Y)(Y - \mu_Y)^2 + \frac{1}{6}f'''(\mu_Y)(Y - \mu_Y)^3 + R_4(Y - \mu_Y)$$

where R_4 is a residual term. Introducing $\bar{Y} = Y - \mu_Y$, $\sigma_Y^2 = \mathbb{V}ar[Y]$ and using results for the moments of a normal distribution (appendix A), we have:

$$\begin{aligned}
\mathbb{E}[f(Y)] \approx & f(\mu_Y) + f(\mu_Y)\mathbb{E}\left[\bar{Y}\right] + \frac{1}{2}f''(\mu_Y)\mathbb{E}\left[\bar{Y}^2\right] + \frac{1}{6}f'''(\mu_Y)\mathbb{E}\left[\bar{Y}^3\right] = f(\mu_Y) + \frac{\sigma_Y^2}{2}f''(\mu_Y) \\
\mathbb{V}ar[f(Y)] \approx & \left(f'(\mu_Y)\right)^2 \mathbb{V}ar\left[\bar{Y}\right] + \frac{\left(f''(\mu_Y)\right)^2}{4}\mathbb{V}ar\left[\bar{Y}^2\right] + \frac{\left(f'''(\mu_Y)\right)^2}{36}\mathbb{V}ar\left[\bar{Y}^3\right] \\
& + f'(\mu_Y)f''(\mu_Y)\mathbb{C}ov\left[\bar{Y}, \bar{Y}^2\right] + \frac{f'(\mu_Y)f'''(\mu_Y)}{3}\mathbb{C}ov\left[\bar{Y}, \bar{Y}^3\right] + \frac{f''(\mu_Y)f'''(\mu_Y)}{6}\mathbb{C}ov\left[\bar{Y}^2, \bar{Y}^3\right] \\
\approx & \left(f'(\mu_Y)\right)^2 \sigma_Y^2 + \frac{\left(f''(\mu_Y)\right)^2}{4}\left(3\sigma_Y^4 - \sigma_Y^4\right) + \frac{\left(f'''(\mu_Y)\right)^2}{36}15\sigma_Y^6 + \frac{f'(\mu_Y)f'''(\mu_Y)}{3}3\sigma_Y^4 \\
\approx & \left(f'(\mu_Y)\right)^2 \sigma_Y^2 + \left(\frac{\left(f''(\mu_Y)\right)^2}{2} + f'(\mu_Y)f'''(\mu_Y)\right)\sigma_Y^4 + \frac{\left(f'''(\mu_Y)\right)^2}{36}15\sigma_Y^6
\end{aligned}$$

and introducing X with mean μ_X, variance σ_X^2, and correlation ρ with Y:

$$\begin{aligned}
\mathbb{C}ov[f(X), f(Y)] \approx & f'(\mu_X)f'(\mu_Y)\mathbb{C}ov[X - \mu_X, Y - \mu_Y] \\
& + \frac{1}{4}f''(\mu_X)f''(\mu_Y)\mathbb{C}ov\left[(X - \mu_X)^2, (Y - \mu_Y)^2\right]
\end{aligned}$$

⚠ these approximations are precise when the higher order moments are small (i.e. mean and variance are small). More precise approximations can be obtained considering higher-order terms:

$$
\begin{aligned}
\mathbb{E}\left[f(Y)\right] &\approx f(\mu_Y) + \frac{\sigma_Y^2}{2} f^{(2)}(\mu_Y) + \frac{\sigma_Y^4}{8} f^{(4)}(\mu_Y) + \frac{\sigma_Y^6}{48} f^{(6)}(\mu_Y) \\
\mathbb{V}ar\left[f(Y)\right] &\approx \left(f^{(1)}(\mu_Y)\right)^2 \sigma_Y^2 + \left(\frac{\left(f^{(2)}(\mu_Y)\right)^2}{2} + f^{(1)}(\mu_Y) f^{(3)}(\mu_Y)\right)\sigma_Y^4 \\
&+ \left(\frac{5\left(f^{(3)}(\mu_Y)\right)^2}{12} + \frac{f^{(2)}(\mu_Y) f^{(4)}(\mu_Y)}{2} + \frac{f^{(1)}(\mu_Y) f^{(5)}(\mu_Y)}{4}\right)\sigma_Y^6 \\
&+ \left(\frac{\left(f^{(4)}(\mu_Y)\right)^2}{6} + \frac{7 f^{(3)}(\mu_Y) f^{(5)}(\mu_Y)}{24}\right)\sigma_Y^8 + \frac{21\left(f^{(5)}(\mu_Y)\right)^2}{320}\sigma_Y^{10}
\end{aligned}
$$

B.2 Application: exponential transformation ($f = \exp$)

Using that $\mathbb{C}ov\left[(X-\mu_X)^2, (Y-\mu_Y)^2\right] \approx 2\rho^2\sigma_X^2\sigma_Y^2$:

$$\mathbb{E}\left[\exp(Y)\right] \approx \exp(\mu_Y)\left(1+\frac{\sigma_Y^2}{2}\right)$$
$$\mathbb{V}ar\left[\exp(Y)\right] \approx \exp(2\mu_Y)\left(\sigma_Y^2+\frac{3}{2}\sigma_Y^4+\frac{15}{36}\sigma_Y^6\right)$$
$$\mathbb{C}ov\left[\exp(X),\exp(Y)\right] \approx \exp(\mu_X+\mu_Y)\left(\rho\sigma_X\sigma_Y+\frac{1}{2}\rho^2\sigma_X^2\sigma_Y^2\right)$$

Note: one can always go one order further to get a better approximation:

$$\mathbb{E}\left[\exp(Y)\right] \approx \exp(\mu_Y)\left(1+\frac{\sigma_Y^2}{2}+\frac{\sigma_Y^4}{8}+\frac{\sigma_Y^6}{48}\right)$$
$$\mathbb{V}ar\left[\exp(Y)\right] \approx \exp(2\mu_Y)\left(\sigma_Y^2+\frac{3}{2}\sigma_Y^4+\frac{7}{6}\sigma_Y^6+\frac{11}{24}\sigma_Y^8+\frac{21}{320}\sigma_Y^{10}\right)$$
$$\mathbb{C}ov\left[\exp(X),\exp(Y)\right] \approx \exp(\mu_X+\mu_Y)\left(\rho\sigma_X\sigma_Y+\frac{1}{2}\rho^2\sigma_X^2\sigma_Y^2 + \frac{1}{2}\rho\left(\sigma_X\sigma_Y^3+\sigma_Y\sigma_X^3\right)+\frac{1}{12}\left(2\rho^3+3\rho\right)\sigma_X^3\sigma_Y^3\right)$$

Illustration: We consider a normally distributed outcome with expectation 1 and variance 0.5 (i.e standard deviation about 0.707). What is its expectation and variance after exp-transformation?

```
set.seed(10); n <- 1e4
mu <- 1; sigma2 <- 0.5

## first order method
mu.exp1 <- exp(mu)
var.exp1 <- exp(2*mu)*sigma2

## third order method
mu.exp2 <- exp(mu)*(1+sigma2/2)
var.exp2 <- exp(2*mu)*(sigma2 + (3/2)*sigma2^2 + (15/36)*sigma2^3)

## n order method
mu.exp3 <- exp(mu)*(1 + sigma2/2 + sigma2^2/8 + sigma2^3/48)
var.exp3 <- exp(2*mu)*(sigma2 + (3/2)*sigma2^2 + (7/6)*sigma2^3 + (11/24)*
    sigma2^4 + (21/320)*sigma2^10)

## empirical value
X.exp <- exp(rnorm(n, mean = mu, sd = sqrt(sigma2)))
mu.expGS <- mean(X.exp)
var.expGS <-  var(X.exp)
```

Comparison mean:

```
rbind(value = c(first.order = mu.exp1,
  second.order = mu.exp2,
  third.order = mu.exp3,
  truth = mu.expGS),
      bias = c(mu.exp1,mu.exp2,mu.exp3,mu.expGS)-mu.expGS,
      relative.bias = (c(mu.exp1,mu.exp2,mu.exp3,mu.expGS)-mu.expGS)/mu.
   expGS)
```

```
              first.order second.order  third.order    truth
value           2.7182818   3.39785229  3.489877452 3.505691
bias           -0.7874091  -0.10783859 -0.015813428 0.000000
relative.bias  -0.2246088  -0.03076101 -0.004510788 0.000000
```

Comparison variance:

```
rbind(value = c(first.order = var.exp1,
  second.order = var.exp2,
  third.order = var.exp3,
  truth = var.expGS),
      bias = c(var.exp1,var.exp2,var.exp3,var.expGS)-var.expGS,
      relative.bias = (c(var.exp1,var.exp2,var.exp3,var.expGS)-var.expGS)/
   var.expGS)
```

```
              first.order second.order third.order    truth
value           3.6945280    6.8502708  7.75513398 8.224438
bias           -4.5299096   -1.3741669 -0.46930364 0.000000
relative.bias  -0.5507865   -0.1670834 -0.05706209 0.000000
```

The second order estimate is much more accurate, especially for the variance.

We now consider a bivariate normally distributed outcome with expectation 0.1, variance 0.1, and correlation 0.5. What is the correlation after exp-transformation?

```
set.seed(10); n <- 1e4
mu <- c(0.1,0.1); sigma2 <- c(0.1,0.1); rho <- 0.5
Sigma <- matrix(c(sigma2[1], rho*sqrt(prod(sigma2)),
    rho*sqrt(prod(sigma2)), sigma2[2]), 2,2)
XY <- mvtnorm::rmvnorm(n, mean = mu, sigma = Sigma)
X <- XY[,1] ; Y <- XY[,2]

cov(exp(X),exp(Y))
exp(mean(X)+2*mean(Y)) * (cor(X,Y)*sd(Y)*sd(X) + 0.5*cor(X,Y)^2*var(Y)*var
    (X))
```

```
[1] 0.06839007
[1] 0.06846545
```

B.3 Application: log-transformation ($f = \log$)

$$\mathbb{E}\left[\log(Y)\right] \approx \log(\mu_Y) - \frac{\sigma_Y^2}{2\mu_Y^2}$$
$$\mathbb{V}ar\left[\log(Y)\right] \approx \frac{\sigma_Y^2}{\mu_Y^2} + \frac{5\sigma_Y^4}{2\mu_Y^4} + \frac{5\sigma_Y^6}{3\mu_Y^6}$$
$$\mathbb{C}ov\left[\log(X), \log(Y)\right] \approx \frac{\rho\sigma_X\sigma_Y}{\mu_X\mu_Y} + \frac{\rho^2\sigma_X^2\sigma_Y^2}{2\mu_X^2\mu_Y^2}$$

Note: one can always go one order further to get a better approximation:

$$\mathbb{E}\left[\log(Y)\right] \approx \log(\mu_Y) - \frac{\sigma_Y^2}{2\mu_Y^2} - \frac{3\sigma_Y^4}{4\mu_Y^4} - \frac{5\sigma_Y^6}{2\mu_Y^6}$$
$$\mathbb{V}ar\left[\log(Y)\right] \approx \frac{\sigma_Y^2}{\mu_Y^2} + \frac{5\sigma_Y^4}{2\mu_Y^4} + \frac{67\sigma_Y^6}{6\mu_Y^6} + \frac{20\sigma_Y^8}{6\mu_Y^8} + \frac{189\sigma_Y^{10}}{5\mu_Y^{10}}$$

Illustration: We consider a normally distributed outcome with expectation 7 and variance 2 (i.e standard deviation about 1.414). What is its expectation and variance after log-transformation?

```
set.seed(10); n <- 1e4
mu <- 7; sigma2 <- 2

## first order method
mu.log1 <- log(mu)
var.log1 <- sigma2/mu^2

## third order method
mu.log2 <-  log(mu) - sigma2/(2*mu^2)
var.log2 <- sigma2/mu^2 + 5*sigma2^2/(2*mu^4) + 5*sigma2^3/(3*mu^6)

## n order method
mu.log3 <-  log(mu) - sigma2/(2*mu^2) - 3*sigma2^2/(4*mu^4) - 5*sigma2^6/
    (2*mu^6)
var.log3 <- sigma2/mu^2 + 5*sigma2^2/(2*mu^4) + 67*sigma2^3/(6*mu^6) + 20*
    sigma2^4/(6*mu^8) + 189*sigma2^5/(5*mu^10)

## empirical value
X.log <- log(rnorm(n, mean = mu, sd = sqrt(sigma2)))
mu.logGS <- mean(X.log)
var.logGS <-  var(X.log)
```

Comparison mean:

```
rbind(value = c(first.order = mu.log1,
  second.order = mu.log2,
  third.order = mu.log3,
  truth = mu.logGS),
      bias = c(mu.log1,mu.log2,mu.log3,mu.logGS)-mu.logGS,
      relative.bias = (c(mu.log1,mu.log2,mu.log3,mu.logGS)-mu.logGS)/mu.
   logGS)
```

```
              first.order second.order  third.order    truth
value          1.94591015 1.9255019858  1.922892529 1.924102
bias           0.02180784 0.0013996795 -0.001209777 0.000000
relative.bias  0.01133403 0.0007274455 -0.000628749 0.000000
```

Comparison variance:

```
rbind(value = c(first.order = var.log1,
  second.order = var.log2,
  third.order = var.log3,
  truth = var.logGS),
      bias = c(var.log1,var.log2,var.log3,var.logGS)-var.logGS,
      relative.bias = (c(var.log1,var.log2,var.log3,var.logGS)-var.logGS)/
   var.logGS)
```

```
               first.order second.order   third.order      truth
value          0.040816327  0.045094589  0.0457541123 0.04632675
bias          -0.005510428 -0.001232166 -0.0005726425 0.00000000
relative.bias -0.118946995 -0.026597277 -0.0123609457 0.00000000
```

The second order estimate is much more accurate, especially for the variance.

B.4 Log-normal distribution

An alternative approach is to use a log-normal distribution. Random variables with log normal distribution have their logarithm equal to a specific value a and their standard deviation equal to a specific value s. So we want to get:

$$\begin{aligned}
\alpha &= \exp(a_0 + \frac{1}{2}s_0^2) \\
\sigma^2 &= \exp(2 * a_0 + s_0^2) * (\exp(s_0^2) - 1) \\
\alpha(1+\gamma) &= \exp(a_1 + \frac{1}{2}s_1^2) \\
\sigma^2 &= \exp(2 * a_1 + s_1^2) * (\exp(s_1^2) - 1)
\end{aligned}$$

So

$$\begin{aligned}
s_0 &= \log\left(1 + \frac{\sigma^2}{\alpha^2}\right) \\
a_0 &= \log(\alpha) - \frac{s_0^2}{2} \\
s_1 &= \log\left(1 + \frac{\sigma^2}{\alpha * (1+\gamma)^2}\right) \\
a_1 &= \log(\alpha * (1+\gamma)) - \frac{s_1^2}{2}
\end{aligned}$$

Illustration: We consider a normally distributed outcome with expectation 7 and variance 2 (i.e standard deviation about 1.414). What is its expectation and variance after log-transformation?

```
set.seed(10); n <- 1e4
X <- rlnorm(1e4, mean=1, sd = 0.5)
## X <- exp(rnorm(1e4, mean=1, sd = sqrt(0.5)))

mu.exp <- mean(X)
sigma2.exp <- var(X)

## taylor expansion method
## mu.exp = exp(mu)*(1 + sigma2/2 + sigma2^2/8 + sigma2^3/48)
## sigma2.exp = exp(2*mu)*(sigma2 + (3/2)*sigma2^2 + (7/6)*sigma2^3 + (11/
    24)*sigma2^4 + (21/320)*sigma2^10)
getSigma2 <- function(sigma2){
    mu.exp^2/sigma2.exp - (1 + sigma2/2 + sigma2^2/8 + sigma2^3/48)^2/(
    sigma2 + (3/2)*sigma2^2 + (7/6)*sigma2^3 + (11/24)*sigma2^4 + (21/320)*
    sigma2^10)
}
var.taylor <- uniroot(f = getSigma2, lower = 1e-5, upper = sigma2.exp)$
    root
mu.taylor <- log(mu.exp/(1 + var.taylor/2 + var.taylor^2/8 + var.taylor^3/
    48))
## mu.taylor <-  log(mu) - sigma2/(2*mu^2) - 3*sigma2^2/(4*mu^4) - 5*
    sigma2^6/(2*mu^6)
## var.taylor <- sigma2/mu^2 + 5*sigma2^2/(2*mu^4) + 67*sigma2^3/(6*mu^6)
    + 20*sigma2^4/(6*mu^8) + 189*sigma2^5/(5*mu^10)

## log distribution method
var.logdist <- log(1+sigma2/mu^2)
mu.logdist <- log(mu) - var.logdist/2

## empirical value
X.log <- log(X)
mu.logGS <- mean(X.log)
var.logGS <-  var(X.log)
```

Comparison mean:

```
rbind(value = c(taylor = mu.taylor,
  dist = mu.logdist,
  truth = mu.logGS),
      bias = c(mu.taylor,mu.logdist,mu.logGS)-mu.logGS,
      relative.bias = (c(mu.taylor,mu.logdist,mu.logGS)-mu.logGS)/mu.logGS
    )
```

```
                     taylor         dist    truth
value           0.999612153  0.998213975 1.000669
bias           -0.001056824 -0.002455001 0.000000
relative.bias -0.001056117 -0.002453360 0.000000
```

Comparison variance:

```
rbind(value = c(taylor = var.taylor,
  dist = var.logdist,
  truth = var.logGS),
      bias = c(var.taylor,var.logdist,var.logGS)-var.logGS,
      relative.bias = (c(var.taylor,var.logdist,var.logGS)-var.logGS)/var.
   logGS)
```

```
                    taylor      dist     truth
value         0.255318149 0.5123473 0.2528091
bias          0.002509088 0.2595382 0.0000000
relative.bias 0.009924835 1.0266175 0.0000000
```