Week #3
January 18-22

1. Delta method

is for approximate calculation of the variance.

Let $X$ be a random variable and $g$ any function. The exact calculation of mean and variance of $g(X)$ involves integral which may be cumbersome. Can we approximate mean and variance?

Yes, using Taylor series expansion.

Denote $E(X) = \mu$ and $var(X) = \sigma^2$.

Mean: Using Taylor series expansion around $x_0 = \mu$ we obtain

\[
g(X) \approx g(\mu) + (X - \mu) \frac{dg}{dx} \bigg|_{x=\mu} = g(\mu) + (X - \mu)g'(\mu). \]

Take expectation:

\[
E(g(X)) \approx E(g(\mu)) + E ((X - \mu)g'(\mu)) = g(\mu) + g'(\mu)E(X - \mu) = g(\mu) + g'(\mu) \cdot 0 = g(\mu)
\]

Thus,

\[
E(g(X)) \approx g(\mu). \tag{1.1}
\]

Variance: Again using Taylor series expansion

\[
g(X) \approx g(\mu) + (X - \mu)g'(\mu),
\]

and using the above,

\[
var(g(X)) = E (g(X) - E(g(X)))^2 \approx E (g(X) - g(\mu))^2
\]
\[
E((X - \mu)g'(\mu))^2 \\
= E((X - \mu)^2g'(\mu))^2 \\
= (g'(\mu))^2E(X - \mu)^2 \\
= \sigma^2(g'(\mu))^2
\]

Thus,
\[
\text{var}(g(X)) \approx \sigma^2(g'(\mu))^2. \tag{1.2}
\]

Prove that (1.1) and (1.2) are exact if \( g \) is a linear function.

**Example.** Let \( X \sim N(\mu, \sigma^2) \). Approximate mean and variance of \( e^X \).

**Solution.** We have
\[
g(x) = e^x, \quad g'(x) = e^x
\]
Therefore,
\[
Ee^X \approx e^\mu, \quad \text{var}(e^X) \approx \sigma^2 e^{2\mu}.
\]

Exact mean can be calculated (\( e^X \) has lognormal distribution),
\[
Ee^X = e^{\mu + .5\sigma^2}.
\]

Delta-method to approximate the mean of the lognormal distribution.

Solid: exact mean (\( \sigma = 1 \))
Dashed line: exact mean (\( \sigma = .5 \))
Dotted line: approximate mean.

For lognormal distribution
\[
\text{var}(e^X) = Ee^{2X} - E^2(e^X) \\
= Ee^Z - E^2(e^X) \\
= Ee^Z - e^{2(\mu+.5\sigma^2)}
\]
where $Z \sim N(2\mu, 4\sigma^2)$. Hence

$$Ee^Z = e^{2\mu + 2\sigma^2}.$$ 

Continue,

$$\text{var}(e^X) = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2} = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$$

Delta-method to approximate the variance of the lognormal distribution.

Solid line: exact variance ($\sigma = .5$)
Dotted line: approximate variance.

2. Normal distribution

is the most important distribution in statistics!

Sometimes it is called Gaussian by the name of German mathematician Carl Gauss.

*Jargon:* A normally distributed RV is called normal RV.

This distribution has two parameters, mean ($\mu$) and variance ($\sigma^2 > 0$). If CRV $X$ has normal distribution we write

$$X \sim N(\mu, \sigma^2).$$

The density of $X$ is

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

and the distribution function is

$$\Phi(x; \mu, \sigma^2) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(t-\mu)^2} dt.$$ 

The density of normally distributed variable is *bell-shaped.*
Three normal densities with unit variance ($\sigma^2 = 1$) and means $\mu = -1$ (dotted), $\mu = 0$ (solid), $\mu = 1$ (dashed).
Parameter $\mu$ determines the location.

Three normal densities with zero mean and different variance:
$\sigma^2 = 1$ (solid), $\sigma^2 = 2$ (dotted), $\sigma^2 = 4$ dotted.
Variance (SD) defines the shape (the width of the bell).

Mean and variance of the normal distribution: if $X \sim N(\mu, \sigma^2)$ then

$$E(X) = \mu \text{ and } var(X) = \sigma^2.$$  

We have for the mean:

$$E(X) = \int_{-\infty}^{\infty} \frac{t}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} \, dt $$

$$= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} (u+\mu) e^{-\frac{1}{2\sigma^2}u^2} \, du $$

$$= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{1}{2\sigma^2}u^2} \, du + \mu \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}u^2} \, du $$

$$= 0 + \mu \times 1 = \mu.$$
Use integration by part to prove $\text{var}(X) = \sigma^2$.

**Linear transformation of a variable with normal distribution leads again to a normal random variable.**

**Proof.** Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$. Then using

$$f_Y(y) = \frac{1}{a} f_X \left( \frac{y - b}{a} \right)$$

we obtain

$$\frac{1}{a} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} (\frac{y-b}{a} - \mu)^2} = \frac{1}{(\sigma a) \sqrt{2\pi}} e^{-\frac{1}{2(\sigma a)^2} (y - a\mu - b)^2}$$

Denote

$$\sigma_Y^2 = (\sigma a)^2 \text{ and } \mu_Y = a\mu + b$$

we see that $Y$ has a normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$.

Fact to remember:

if $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ then $Y \sim N(a\mu + b, a^2\sigma^2)$.

Standard normal distribution $N(0, 1)$. If $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma}$$

is called $Z$-score or standardized normal random variable, and

$$Z \sim N(0, 1)$$

i.e. it has standard normal distribution.

The standard normal density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} x^2}$$

and the standard normal distribution is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2} t^2} dt.$$  

Tables are provided in the Appendix A (Rice, p.A7).

It implies that if $X \sim N(\mu, \sigma^2)$ then

$$\Pr(X \leq a) = F_X(a) = \Phi \left( \frac{a - \mu}{\sigma} \right).$$

**Problem 1.** Let $X \sim N(0, 1)$, find the median, and the lower and upper quartiles.

**Solution.** Since $X$ is symmetric, median=mode=mean=0. Also since $X$ is symmetric lower quartile=upper quartile. From Table 2 we find that approximately the upper quartile, $x_{.75} = .57$, and $x_{.25} = -.57$. Therefore, $\Pr(-.57 < X < .57) = .5$.

**Problem 2.** Let $X \sim N(1, 4)$, find $\Pr(|X| < 2)$.
Density of $X \sim N(1, 4)$.

**Solution.** We have

\[
\Pr(|X| < 2) = \Pr(-2 < X < 2) = \Pr(X < 2) - \Pr(X < -2) = \Phi\left(\frac{2-1}{2}\right) - \Phi\left(\frac{-2-1}{2}\right) = \Phi(0.5) - \Phi(-1.5)
\]

Now look up Table 2 (Rice, A7).

\[
\Phi(0.5) = .6915, \\
\Phi(-1.5) = 1 - \Phi(1.5) = 1 - .9332 = .0668
\]

Finally,

\[
\Pr(|X| < 2) = .6915 - .0668 = .6247.
\]

**Problem 3.** Given $X \sim N(\mu, \sigma^2)$, find the distribution and the density function of $Y = X^2$.

**Solution.** Warning: we cannot use the formula for the distribution and density derived above because $g(x) = x^2$ is not monotone. In order to find the distribution of $Y$ we proceed directly from the definition of the distribution function (sometimes this method is called "method of distribution function"):

\[
F_Y(y) = \Pr(Y \leq y) = \Pr(X^2 \leq y) = \Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = \Pr(X \leq \sqrt{y}) - \Pr(X < -\sqrt{y}).
\]

where $y > 0$. Recall, if $X \sim N(\mu, \sigma^2)$

\[
\Pr(X < a) = \Phi\left(\frac{a - \mu}{\sigma}\right)
\]
where \( \Phi \) is the standard normal distribution. Therefore,

\[
\Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi \left( \frac{\sqrt{y} - \mu}{\sigma} \right) - \Phi \left( \frac{-\sqrt{y} - \mu}{\sigma} \right)
\]

and the distribution function of \( Y \) is

\[
\Pr(Y \leq y) = \Phi \left( \frac{\sqrt{y} - \mu}{\sigma} \right) - \Phi \left( \frac{-\sqrt{y} - \mu}{\sigma} \right)
\]

Take derivative with respect to \( y \) and obtain the density (use \( \Phi' = \phi \) and the chain rule)

\[
\frac{1}{2\sigma \sqrt{y}} \phi \left( \frac{\sqrt{y} - \mu}{\sigma} \right) + \frac{1}{2\sigma \sqrt{y}} \phi \left( \frac{-\sqrt{y} - \mu}{\sigma} \right)
\]

\[
= \frac{1}{2\sigma \sqrt{y}} \left[ \phi \left( \frac{\sqrt{y} - \mu}{\sigma} \right) + \phi \left( \frac{-\sqrt{y} - \mu}{\sigma} \right) \right]
\]

If \( \mu = 0 \) then the density is simpler

\[
\frac{1}{\sigma \sqrt{y}} \phi \left( \frac{\sqrt{y}}{\sigma} \right)
\]

or

\[
\frac{1}{\sigma \sqrt{2\pi y}} e^{-\frac{1}{2} \left( \frac{\sqrt{y}}{\sigma} \right)^2} = \frac{1}{\sigma \sqrt{2\pi y}} e^{-\frac{y}{2\sigma^2}}
\]

Density function of \( X^2 \) where \( X \sim N(0, 1) \).

**Problem** (continued). You need to measure the area of the square. You measured the side of the square, 10 feet with the measurement error SD=1. Assuming the measurement follows normal distribution, what is the probability that the area of the square is more than:

(a) 100 sq. feet,
(b) \((10+1)^2 = 121\) sq. feet?

**Solution.** Let \( X \) be the length of the square side. We know that \( X \sim N(\mu, \sigma^2) \) where \( \mu = 10, \sigma = 1 \). The first probability is

\[
\Pr(X^2 > 100) = 1 - \Pr(X^2 \leq 100) = 1 - F_X(100)
\]
where \( Y = X^2 \). But

\[
F_Y(100) = \Phi \left( \frac{\sqrt{100} - \mu}{\sigma} \right) - \Phi \left( \frac{-\sqrt{100} - \mu}{\sigma} \right)
\]

\[
= \Phi \left( \frac{\sqrt{100} - 10}{1} \right) - \Phi \left( \frac{-\sqrt{100} - 10}{1} \right)
\]

\[
= \Phi (0) - \Phi (-20)
\]

\[
\simeq \Phi (0)
\]

\[
= \frac{1}{2}.
\]

Finally,

\[
1 - F_Y(100) = 1 - \frac{1}{2} = \frac{1}{2}.
\]

The second probability is

\[
\Pr(X^2 > 121) = 1 - F_Y(121).
\]

But

\[
F_Y(121) = \Phi \left( \frac{\sqrt{121} - 10}{1} \right) - \Phi \left( \frac{-\sqrt{121} - 10}{1} \right)
\]

\[
= \Phi (1) - \Phi (-21)
\]

\[
\simeq \Phi (1)
\]

\[
= .8413.
\]

Finally,

\[
\Pr(X^2 > 121) \simeq 1 - .8413
\]

\[
= .1587
\]

What assumption is week?

What is unrealistic in this problem?

How would you reformulate this problem to make it more realistic?

**Problem 4.** \( X \) is normally distributed with mean=\( \mu \) and variance \( \sigma^2 \). Find the probabilities

\[
\Pr(|X - \mu| < 2\sigma) \text{ and } \Pr(|X - \mu| < 3\sigma).
\]

**Solution.**

\[
\Pr(|X - \mu| < 2\sigma)
\]

\[
= 1 - 2 \cdot \Phi (-2) = 0.9544997
\]

\[
\simeq 0.954
\]
The rule of 2 sigma:
if RV is normally distributed then with probability
\[ \Pr(|X - \mu| < 3\sigma) = 1 - 2 \cdot \Phi(-3) = 0.9973002 \approx 0.997 \]
the range of RV is plus/minus 2 sigma around the mean.

The rule of 3 sigma:
if RV is normally distributed then with probability close to 1 the range of RV is
plus/minus 3 sigma around the mean.

3. Homework (due Wednesday, January 27)

Maximum number of points is 29.

1. (5 points). During summer John decided to get some extra cash by painting. John had to
paint a square wall. He roughly measured the width=10 (mean) with the error plus/minus 2 feet
(SD). Assuming that his measurement has normal distribution and he spends $\frac{1}{4}$ gallon/sq. foot of
paint, what is the expected amount of paint he needs to buy? What is the probability to run out of
paint if he buys 25 gallons? How many gallons is needed to buy to run out of paint with probability
0.95?

Solution. Let \( X \) denote the size, we know that \( X \sim N(10, 4) \). Then, the expected area is
\[ E(X^2) = \mu^2 + \sigma^2 = 100 + 4 = 104 \]
Thus, he needs \( 104 \times 0.25 = 26 \) gallons of paint. If he buys 25 gallons than the probability to run out of paint is
\[
\Pr(X^2 > 25 \times 4) = \Pr(X^2 > 100) = \Pr(|X| > 10) = \Pr(X > 10) + \Pr(X < -10) = 1 - \Phi\left(\frac{10 - 10}{2}\right) + \Phi\left(-\frac{10 - 10}{2}\right)
\]
\[ = 1 - \Phi(0) + \Phi(-10) \approx \frac{1}{2} \]
Let \( g \) be the amount of paint. Then John runs out of paint with probability 0.95 if \( \Pr(X^2 > 4g) = 0.95 \). Since
\[
\Pr(X^2 > 4g) = 1 - \Phi\left(\frac{2\sqrt{g} - 10}{2}\right) + \Phi\left(-\frac{2\sqrt{g} - 10}{2}\right)
\]
The quantity $\Phi \left( \frac{-2\sqrt{g} - 10}{2} \right)$ is negligible, so that the needed probability can be approximated as

$$\Phi \left( \frac{2\sqrt{g} - 10}{2} \right) = .05$$

which gives

$$\frac{2\sqrt{g} - 10}{2} = -1.65,$$

and the amount of paint is $g = (5 - 1.65)^2 = 11.2$ gallons.

3. (2 points). Let $E(X) = \mu$ and $var(X) = \sigma^2$. Show that for RV $Z = (X - \mu)/\sigma$ we have $E(Z) = 0$ and $var(Z) = 1$.

Solution. We have

$$E(Z) = E \left( \frac{X - \mu}{\sigma} \right) = \frac{1}{\sigma} E(X - \mu) = \frac{1}{\sigma} (E(X) - \mu) = \frac{1}{\sigma} (\mu - \mu) = 0$$

$$var(Z) = var \left( \frac{X - \mu}{\sigma} \right) = \frac{1}{\sigma^2} var(X - \mu) = \frac{1}{\sigma^2} var(X) = \frac{1}{\sigma^2} \sigma^2 = 1.$$ 

4. (3 points). Find the density of $|X|$ where $X \sim N(\mu, \sigma^2)$. Approximate the probability $Pr(|X| > c)$ for large $\mu$ and relatively small $\sigma$.

Solution. First we find the distribution of $|X|:

$$F_{|X|}(x) = Pr(|X| \leq x) = Pr(X \leq x) - Pr(X \leq -x)$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx - \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{-c} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx.$$ 

Then, the density is the derivative of $F_{|X|}(x)$,

$$f_{|X|}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} + \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x+\mu)^2}$$

$$= \frac{1}{\sigma \sqrt{2\pi}} \left( e^{-\frac{1}{2\sigma^2}(x-\mu)^2} + e^{-\frac{1}{2\sigma^2}(x+\mu)^2} \right).$$

The probability $Pr(|X| > c)$ is

$$1 - Pr(X \leq c) + Pr(X \leq -c) = 1 - \Phi \left( \frac{c - \mu}{\sigma} \right) + \Phi \left( \frac{-c - \mu}{\sigma} \right).$$

Since $\mu$ is large, the amount $\frac{-c - \mu}{\sigma}$ is negative and very large in absolute value. This implies that

$$\Phi \left( \frac{-c - \mu}{\sigma} \right) \approx 0$$

so that

$$Pr(|X| > c) \approx 1 - \Phi \left( \frac{c - \mu}{\sigma} \right) = Pr(X > c).$$

5. (6 points). The lognormal distribution is defined as the distribution of $e^X$ where $X \sim N(\mu, \sigma^2)$. Derive the density of this distribution. Find the mean of $e^X$. 

10
Solution. We use formula

\[ f_Y(y) = \frac{dg^{-1}(y)}{dy} f_X(g^{-1}(y)) \]

where \( g = \exp \) and \( g^{-1} = \ln \). Since

\[ f_X(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \]

and

\[ \frac{dg^{-1}(y)}{dy} = \frac{1}{y} \]

we obtain for \( Y = \exp(X) \)

\[ f_Y(y) = \frac{1}{y \sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln y - \mu)^2}, \quad y > 0. \]

To find the mean we calculate the integral

\[ E(e^X) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^x e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \, dx = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{x-\frac{1}{2\sigma^2}(x-\mu)^2} \, dx. \]

But

\[ x - \frac{1}{2\sigma^2}(x-\mu)^2 \]
\[ = -\frac{1}{2\sigma^2} \left( (x-\mu)^2 - 2\sigma^2 x \right) = -\frac{1}{2\sigma^2} \left( x^2 - 2x\mu + \mu^2 - 2\sigma^2 x \right) \]
\[ = -\frac{1}{2\sigma^2} \left( x^2 - 2x\mu + \mu^2 - 2\sigma^2 x \right) \]
\[ = -\frac{1}{2\sigma^2} \left( x - (\mu + \sigma^2)^2 \right) \]
\[ = -\frac{1}{2\sigma^2} \left( \mu^2 - (\mu + \sigma^2)^2 \right) \]
\[ = -\frac{1}{2\sigma^2}(x - (\mu + \sigma^2))^2 - \frac{1}{2\sigma^2}(\mu^2 - (\mu + \sigma^2)^2). \]

But

\[ \frac{1}{2\sigma^2}(\mu^2 - (\mu + \sigma^2)^2) = \frac{1}{2\sigma^2}(-2\sigma^2 \mu + \sigma^4) = -\mu - 5\sigma^2 \]

Then,

\[ E(e^X) = e^{\mu + 5\sigma^2} \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-(\mu+\sigma^2))^2} \, dx \]
\[ = e^{\mu + 5\sigma^2}. \]

6. (4 points) Scores on an examination are assumed to be normally distributed with a mean of 78 and a variance 36.

(a) What is the probability that a person taking the examination scores higher than 72?.
(b) Suppose that students scoring in the top 10% of this distribution are to receive an A grade. What is the minimum score a student must achieve to earn an A grade?

Solution. Let $X$ be the score, we know that $X \sim N(78, 36)$. The first probability is

$$\Pr(X > 72) = 1 - \Pr(X \leq 72) = 1 - \Phi \left( \frac{72 - 78}{\sqrt{36}} \right) = 1 - \Phi \left( \frac{72 - 78}{6} \right) = 1 - \Phi (-1) = 1 - .1587 = .9413.$$ 

The minimum score, $x$ is the solution to the following equation

$$\Pr(X > x) = .1$$

or

$$\Pr(X \leq x) = \Phi \left( \frac{x - 78}{6} \right) = .9$$

If $x_{.9}$ is the 90% quantile then

$$\frac{x - 78}{6} = x_{.9}$$

But from Rice, Table 2, $x_{.9} = 1.28$ we obtain

$$\frac{x - 78}{6} = 1.28$$

which gives $x = 78 + 1.28 \times 6 = 85.68$.

7. (4 points). Does the rule of 3 sigma works for uniform distribution? What is the probability to fall in the range of 3 sigma around the mean or a uniform distribution?

Solution. Let $X \sim U(a, b)$. Then we know

$$E(X) = \frac{a + b}{2}, \quad var(X) = \frac{(b - a)^2}{12}.$$ 

Also, we know that for $X \sim U(a, b)$

$$\Pr(c < X < d) = \frac{d - c}{b - a}.$$ 

In our case

$$d = \frac{a + b}{2} + 3 \frac{b - a}{2\sqrt{3}}, \quad c = \frac{a + b}{2} - 3 \frac{b - a}{2\sqrt{3}}$$

and

$$d - c = 6 \frac{b - a}{2\sqrt{3}} = \sqrt{3}(b - a)$$

and

$$\frac{d - c}{b - a} = \frac{\sqrt{3}(b - a)}{b - a} = \sqrt{3} > 1.$$ 

This means that the interval $(c, d)$ cover the interval $(a, b)$. In other words, for uniform RV $\pm 3\sigma$ covers the entire range of the variable: the rule of 3 sigma works.
8. (5 points). The US standard for fat concentration in whole milk is from 11.5 to 12.5 percent. A Vermont company produces milk with the mean fat concentration 12.2 percent and SD=.2. Assuming that the fat concentration has normal distribution, what is the probability that the fat concentration in a pack of milk taken at random: (a) will be more than 12.5, (b) will not meet the US standard?

Solution. If \( X \) denotes the fat concentration in Vermont milk then we know that it varies according \( X \sim N(12.2, 0.2^2) \). The first probability is

\[
\Pr(X > 12.5) = 1 - \Pr(X \leq 12.5) = 1 - \Phi \left( \frac{12.5 - 12.2}{0.2} \right)
\]

\[
= 1 - \Phi (1.5) = 1 - .9332 = 0.0668.
\]

The second probability is

\[
\Pr(X > 12.5 \text{ OR } X < 11.5) = \Pr(X > 12.5) + \Pr(X < 11.5)
\]

\[
= 0.0668 + \Phi \left( \frac{11.5 - 12.2}{0.2} \right)
\]

\[
= 0.0668 + \Phi (-3.5) =
\]

\[
= 0.0668 + (1 - \Phi (3.5))
\]

\[
= 0.0668 + (1 - 0.9999)
\]

\[
= 0.0669
\]

\[
\approx 0.0668.
\]