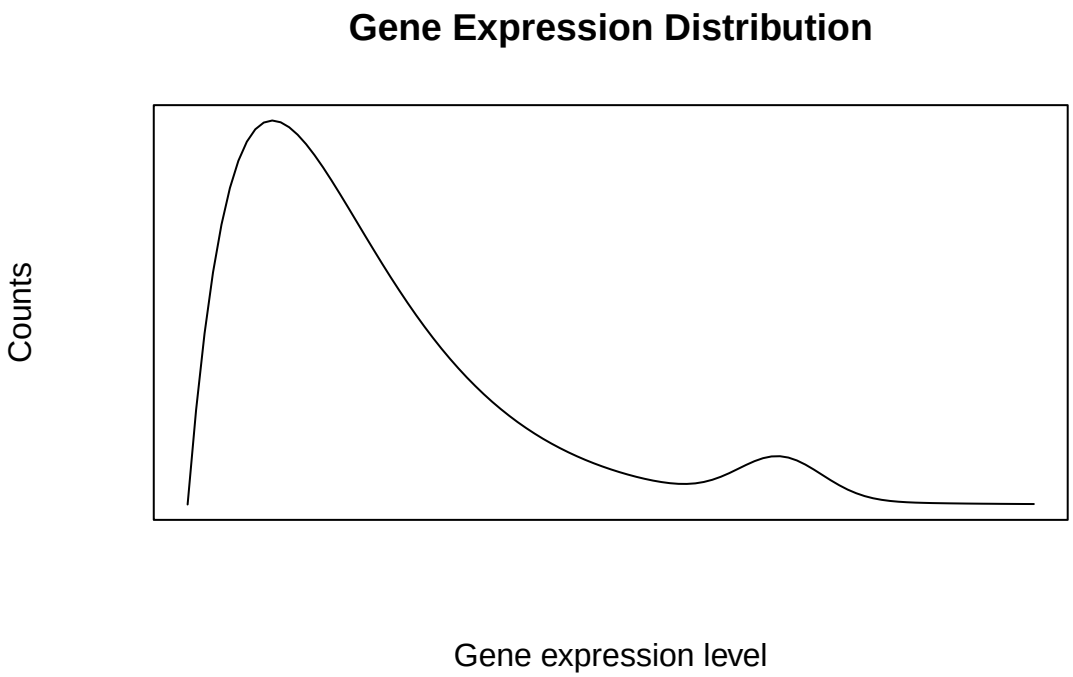
Biostatistics Homework 4

By 𝕃umi (张鹿鸣 12112618)

1 Gene Expression Distribution

1.1)

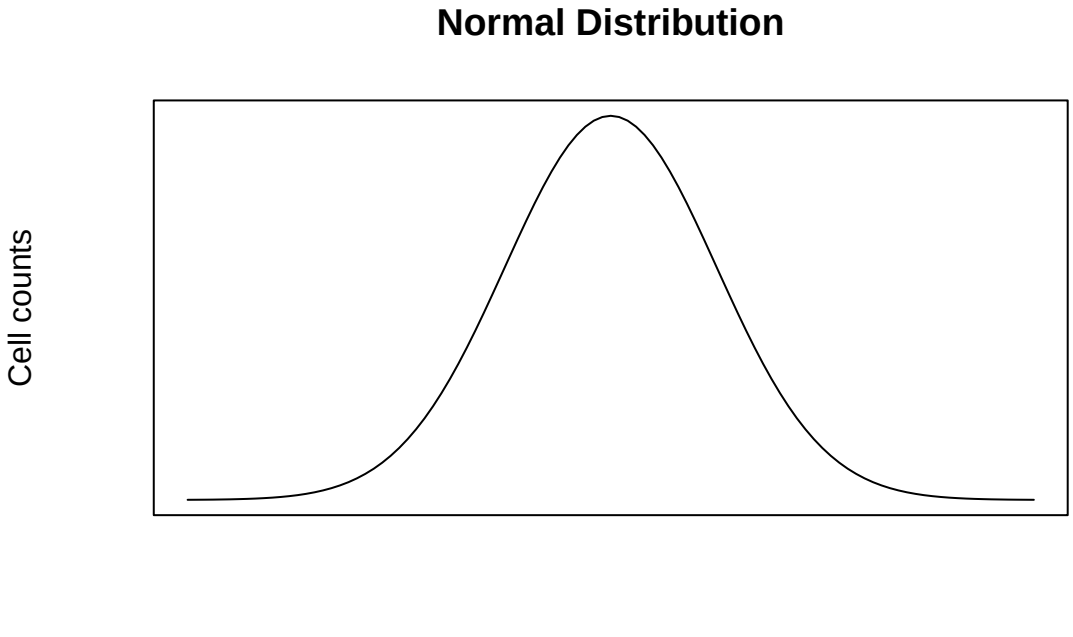
```
curve(dgamma(x, shape = 2, scale = 1) + dnorm(x, mean = 7, sd = 0.5) / 20,
      from = 0, to = 10,
      main = "Gene Expression Distribution",
      xlab = "Gene expression level", ylab = "Counts",
      xaxt = 'n', yaxt = 'n')
```



I think the gene expression pattern will follow a Gamma distribution, which in other words is a distribution skewed to the right. Because many genes are not significantly expressed, thus most of the genes will have a rather low expression level. But some genes are expressed in a very significant amount, because they might be essential for cell survival, or are the signatures for this specific cell type. So a few counts will lie in the high expression area. Thus I think a Gamma distribution with a small bump on the right will fit best.

1.2)

```
curve(dnorm(x, mean = 0, sd = 1),
      from = -4, to = 4,
      main = "Normal Distribution",
      xlab = "Gene expression level", ylab = "Cell counts",
      xaxt = 'n', yaxt = 'n')
```



I think the *Nanog* gene expression across the embryonic stem cell population should follow a normal distribution. Because they are all from the same embryonic stem cell population, their expression of a particular gene should not vary from a certain value vary dramatically, although there might be fluctuations.

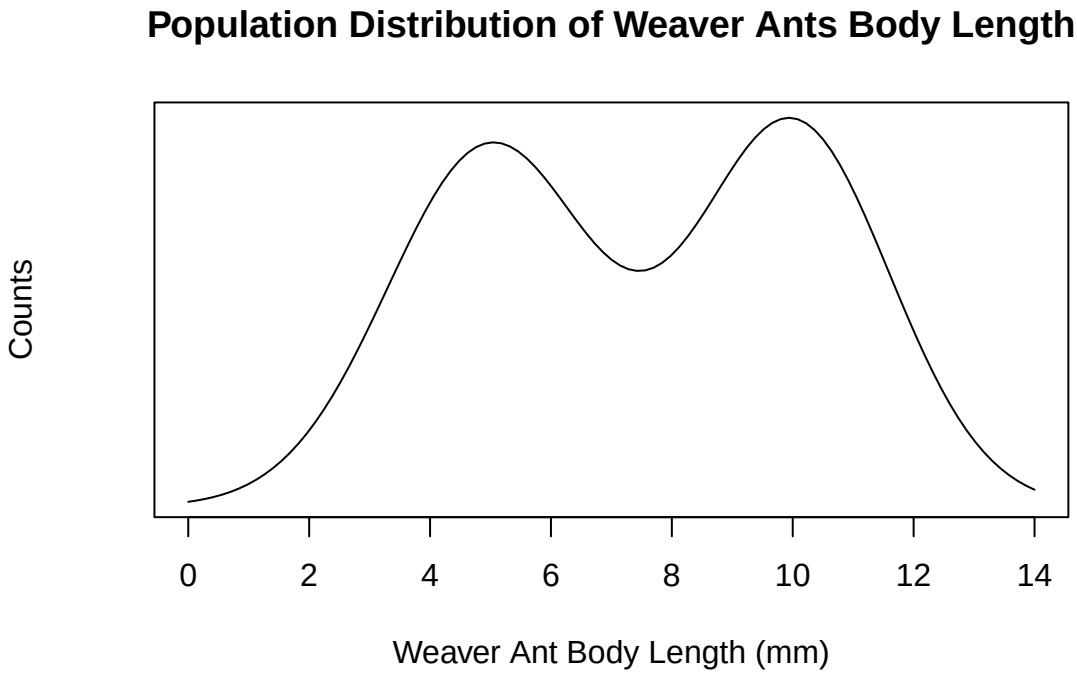
1.3)

I think the distributions of 1.1) and 1.2) are not the same. 1.1) is the distribution of expression levels of all the genes in a single cell, it shows that genes are expressed differently in this cell. 1.2) is the distribution of the expression of a gene in multiple cells, it shows the general level of expression of this gene in all the cells.

2. Weaver ants

2.1)

```
curve(dnorm(x, mean = 5, sd = 1.7) + dnorm(x, mean = 10, sd = 1.6),
      from = 0, to = 14,
      main = 'Population Distribution of Weaver Ants Body Length',
      xlab = 'Weaver Ant Body Length (mm)',
      yaxt = 'n', ylab = 'Counts')
```



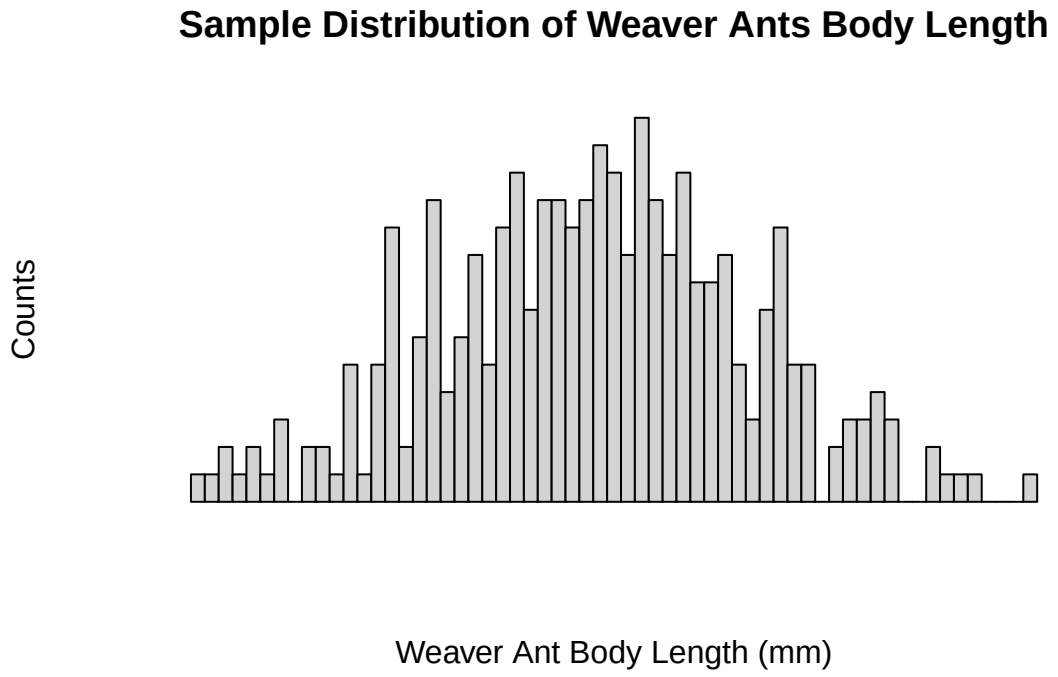
Because there are two types of weaver ants with different body lengths, I suspect that the curve can have two peaks accounting for the two different types of ants. Among the two different types of ants, the body length should follow a normal distribution respectively, so the population distribution of the body length should simply be the addition of the two normal distributions.

2.2)

Since the sample is drawn randomly from the population, we can assume that the sample follows the same type of distribution of the population.

```
hist(rnorm(312, mean = 5, sd = 1.7) + rnorm(313, mean = 10, sd = 1.6),
     main = 'Sample Distribution of Weaver Ants Body Length',
     xlab = 'Weaver Ant Body Length (mm)', xaxt = 'n',
     yaxt = 'n', ylab = 'Counts',
     breaks = 50)
```

```
## Warning in rnorm(312, mean = 5, sd = 1.7) + rnorm(313, mean = 10, sd = 1.6):
## longer object length is not a multiple of shorter object length
```



2.3)

When you randomly choose multiple groups of 625 weaver ants, the mean body length should follow a normal distribution with $\mu = 7, \sigma^2 = (\frac{10}{25})^2$. Suppose X_i is the random variable indicating the mean body length for the 625 weaver ants.

$$P(X_i \leq 7.5) = P(Z \leq \frac{7.5 - \mu}{\sigma}) = P(Z \leq 1.25) \approx 0.8944$$

So the probability of getting a mean from 625 weaver ants that is smaller than 7.5mm is 0.8944.

3 A Simple Random Sample

C

4 The Concept of Population

[tick] It is better to think of our population as an abstract concept.

[tick] The population is effectively infinite.

5 A Practical Simple Random Sampling Strategy

```
sample(1:1000, 20)
```

```
##  [1] 543 834 491 485 269 681 249 178 247 212  31 910 798 516 902 476 890 972 815
## [20] 494
```

But actually, I think just taking the first 20 participants is OK, if your 1000 participants are labeled randomly.

6 Triceps skinfold thickness

6.1)

Similar to problem 2 about the ants, suppose the mean triceps skinfold thickness of 25 people follows is the random variable X. Than X should follow a normal distribution with parameters $\mu = 1.35, \sigma^2 = 0.01$.

$$P(X \leq 1.05) = P(Z \leq \frac{1.05 - \mu}{0.1}) = P(Z \leq -3) \approx 0.0013$$

6.2)

From 6.1) we can see that, getting 25 random people and observing a mean value smaller than 1.05 is very rare ($p < 0.0013$). From the possibility, you can say that it's a once in a thousand opportunity to get a random sample like this. We can also calculate the possibility to observe a mean larger than \$X = 0.92\$, which is $p = 0.99999$. It is highly unlikely to get a value smaller than 0.92cm. Thus we can say that these people who suffer from COPD are significantly different from the normal population.

7 Decrease Manufacturing Variability

7.1)

Suppose that the weights of bags of potato chips follows the random variable X.

$$P(X \leq 330) = P(Z \leq \frac{330 - 362}{20}) = P(Z \leq -1.6) \approx 0.0548$$

7.2)

Suppose the random variable Y describes the mean of the sample with size 100. Y follows a normal distribution $Y \sim N(362, 4)$.

$$P(360 < Y < 363) = P(-1 < Z < 0.5) = 0.5328$$

7.3)

Because only 1% of bags weigh less than 330g, the Z score for 1% is -2.326.

$$\frac{330 - 362}{\sigma} = -2.326$$
$$\sigma \approx 13.75$$

8 A Hypothetical Linear PDF

8.1)

$$\int_0^1 (1 - \theta) + 2\theta x dx = [(1 - \theta)x + \theta x^2]_0^1 = 1$$

This shows that the area under the PDF is 1.

$$0 < x < 1$$
$$1 - \theta < f_{\mathbf{X}}(x) < 1 + \theta$$

Because $-1 < \theta < 1$, $0 < f_{\mathbf{X}}(x) < 2$.

So this probability density functions fulfills the basic properties of the PDF. That is, the area under the curve is 1, and all the density values are larger than 0.

8.2)

$$F_X(x) = \int_0^x (1-\theta) + 2\theta t dt = (1-\theta)x + \theta x^2$$

So

$$F_X(x) = \begin{cases} 1 & , x > 1 \\ (1-\theta)x + \theta x^2 & , 1 > x > 0 \\ 0 & , x \leq 0 \end{cases}$$

8.3)

$$\begin{aligned} E(X) &= \int_0^1 x(1-\theta) + 2\theta x^2 dx \\ &= \frac{(1-\theta)x^2}{2} + \frac{2\theta x^3}{3}]_0^1 \\ &= \frac{3+\theta}{6} \end{aligned}$$

8.4)

$$\mathcal{L}(\theta; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} [(1-\theta) + 2\theta x_i]$$

So, $\ell(\theta; x) = \ln \mathcal{L}(\theta; x)$

$$\ell(\theta; x) = \sum_{i=1}^{n} \ln [(1-\theta) + 2\theta x_i]$$

8.5)

$$\frac{d\ell}{d\theta} = \sum_{i=1}^{n} \frac{2x_i - 1}{(1-\theta) + 2\theta x_i} = 0$$

$$\hat{\theta} = t$$

$$\sum_{i=1}^{n} \frac{2x_i - 1}{(1-\hat{\theta}) + 2\hat{\theta} x_i} = \sum_{i=1}^{n} \frac{2x_i - 1}{(1-t) + 2t x_i} = 0$$

So we choose B.

9 Waiting Time Between Text Messages

$$\begin{aligned}
\mathcal{L}(\lambda; x) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \\
\ell(\lambda; x) &= \ln \mathcal{L}(\lambda; x) \\
&= n \ln \lambda - \lambda \sum_{i=1}^{n} x_i \\
\frac{d\ell}{d\lambda} &= 0 = \frac{n}{\widehat{\lambda}} - \sum_{i=1}^{n} x_i \\
\widehat{\lambda} &= \frac{n}{\sum_{i=1}^{n} x_i}
\end{aligned}$$

So the maximum likelihood estimation of λ is $\frac{n}{\sum_{i=1}^{n} x_i}$.