Probability

February 5, 2013 Debdeep Pati

Estimation

1. Statistical problems - a) Distribution is known. b) Distribution is unknown.
2. When Distribution is known, then we can have either i) Parameters are known or ii) Parameters of the distribution are unknown
3. Estimation: Want to estimate the values of specific population parameters
4. Hypothesis testing: Testing whether the value of a population parameter is equal to some specific value.

Estimation problems

1. Measurements of systolic blood pressures of a group of people, which are believed be follow normal distribution. How can we estimate the parameters (μ, σ^2)?
2. Estimation of the prevalence of HIV-positive people in a low-income community - If we assume the number of cases among n people sampled is binomial with parameter p, how is the parameter p estimated?
3. Interested in both Point estimation and Interval estimation

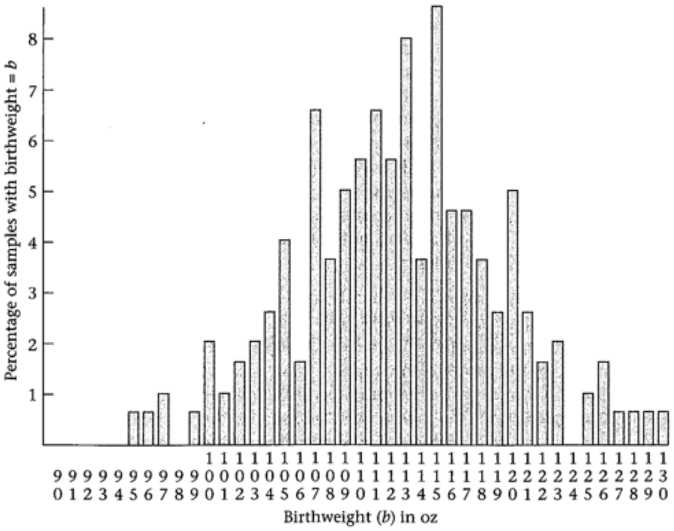
Estimation of the Mean of a Distribution

1. A natural estimation of the population mean μ is the sample mean

$$\bar{X} = \sum_{i=1}^{n} X_i$$

2. Since each X_i's are assumed to be random variables, the quantity $\bar{X}$ is also random.
3. Let $X_1, \ldots, X_n$ be a random sample drawn from some population with mean μ. Then $E(\bar{X}) = \mu$.
4. An estimator $\hat{\theta}$ of a parameter θ is unbiased $E(\hat{\theta}) = \theta$.
5. $\bar{X}$ is the minimum variance unbiased estimator of μ.

Sampling distribution of $\bar{x}$ over 200 samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0–100.9, etc.)



6. Variance of the mean: $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \sigma^2/n$ assuming $V(X_i) = \sigma^2$ for all i.

7. Standard Error of the mean: Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with underlying mean μ and variance σ^2. The set of sample means in repeated random samples of size n from this population has variance σ^2/n. The standard deviation of this set of sample means is $\sigma/\sqrt{n}$ and is referred to as the standard error of the mean (sem) of the standard error.

   The standard error of the mean, or the standard error, is given by $\sigma/\sqrt{n}$ and is estimated by $s/\sqrt{n}$. The standard error represents the estimated standard deviation obtained form a set of sample means from repeated samples of size n from a ovulation with underlying variance σ^2.

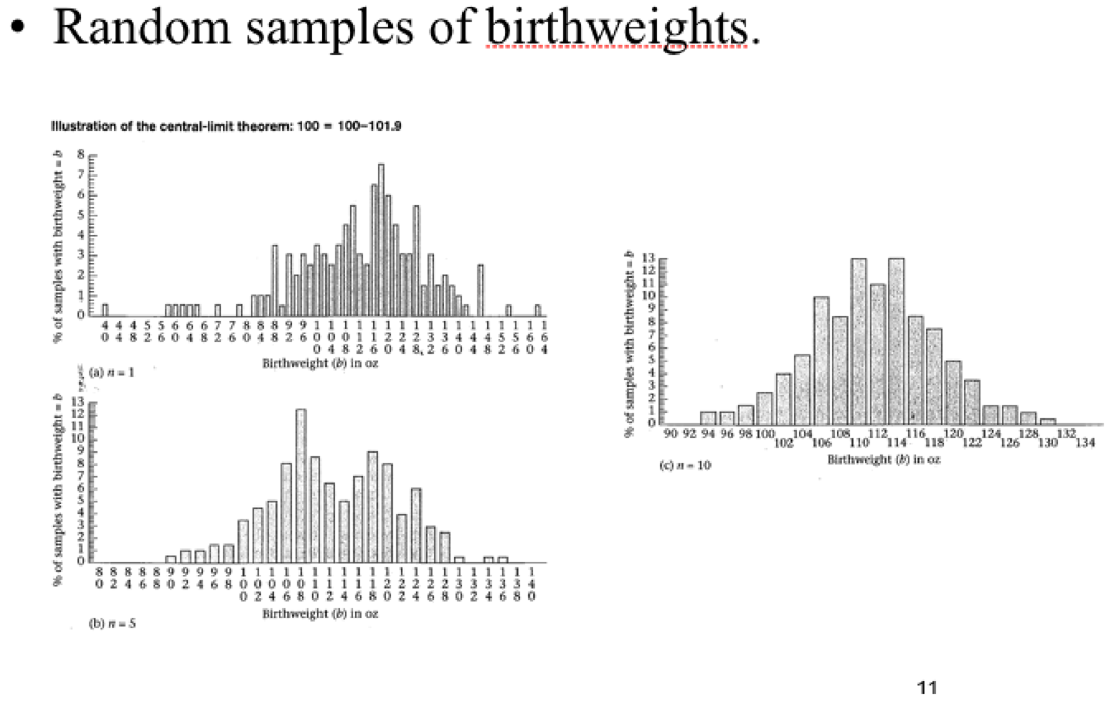
8. Ex. Compute the standard error of the mean for the following sample of birth weights. $97, 125, 62, 120, 132, 135, 118, 137, 126, 118$.

$$\bar{X} = \sum_{i=1}^{n} X_i/n, \quad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

9. The mean is 117 and standard deviation is 22.4. Hence $s/\sqrt{n} = 22.44/\sqrt{10} = 7.09$.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a ranom sample from a population with underlying mean μ and variance σ^2, then $\bar{X} \sim N(\mu, \sigma^2/n)$. Many distributions encountered in practice are not normal, but sampling distribution of the sample average is approximately normal.



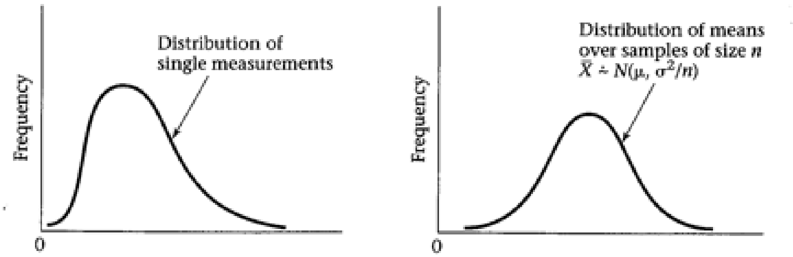
- Random samples of birthweights.

Serum triglyceride distribution tends to be positively skewed, with a few people with very high values. The mean over samples of size n is normally distributed

Sampling distribution

Suppose that we draw all possible samples of size n from a given population. Suppose further that we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. The probability distribution of this statistic is called a sampling distribution.
Example(Obstetrics): Compute the probability that the mean birthweight from a sample of 10 drawn from 1000 infants will fall between 98.0 and 126.0 oz. the mean birthweight for the 1000 birthweights is 112.0 and standard deviation is 20.6. Assuming $\bar{X}$ follows a normal distribution with mean $\mu = 112oz$ and standard deviation $\sigma/\sqrt{n} = 20.6\sqrt{10} = 6.51$.

Figure 6.5 **Distribution of single serum-triglyceride measurements and of means of such measurements over samples of size *n***



(a) Individual serum-triglyceride values

(b) Mean serum triglycerides

12

Then we need to calculate

$$P(98.0 \leq \bar{X} \leq 126.0) = \Phi(\frac{126.0 - 112.0}{6.51}) - \Phi(\frac{98.0 - 112.0}{6.51}) = 0.968$$

We can also do this in R by typing

```
pnorm(126, 112, 20.6) - pnorm(98,112, 20.6)
```

1 Interval Estimation

1. Quantify the uncertainty
2. The 10 birthweigths $97, 125, 62, 120, 132, 135, 118, 137, 126, 118$ have a mean of 116.9 oz. How certain are we that the true mean is 116.9 oz? $116.9oz \pm 1$ oz and $116.9oz \pm 1$ lb are certainly different.
3. The sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$. If μ and σ^2 are known then if you keep on generating samples, 95% of all such samples will fall in the interval
$$(\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n})$$
4. We can also express the mean in standardized form by
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

5. 95% of Z value from repeated samples of size n will fall between -1.96 and +1.96.

6. However, the assumption that σ is known is quite artificial. Since σ is unknown, we can estimate σ by the sample standard deviation s and construct confidence intervals using

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

7. This quantity is no longer normally distributed. The distribution is called Students t distribution, or t distribution if X_i's are normally distributed. t distribution is not a unique distribution. It is a family of distributions indexed by a parameter, the degrees of freedom (df).

8. If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and are independent, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is distributed as a t distribution with $n-1$ degrees of freedom.

9. The $100 \times u$ th percentile of a t distribution is d degrees of freedom is denoted by $t_{d,u}$, that is,

$$P(t_d < t_{d,u}) = u$$

10. $t_{20,0.95}$ stands for the95 th percentile or the upper 5th percentile of a t distribution with 20 degrees of freedom.

1.1 Comparison of t and Normal distribution

1. Compare t with d degrees of freedom to N(0, 1). For any $\alpha > 0.5, t_{d,1-\alpha}$ is always larger than the corresponding percentile for an $N(0,1)(z_{1-\alpha})$. When d becomes large t converge to $N(0,1)$.

2. Find the upper 5th percentile of a t distribution with 23 df. $t_{23,.95}$ is given in row 23 and column 0.95 of Table 5 and is 1.714. Probabilities associated with t distribution can also be calculated using statistical programs. In $R : qt(.95, 23)$

3. If σ is unknown we want $100(1-\alpha)\%$ of the t statistics should fall between the lower and upper quantile of a t_{n-1} distribution.

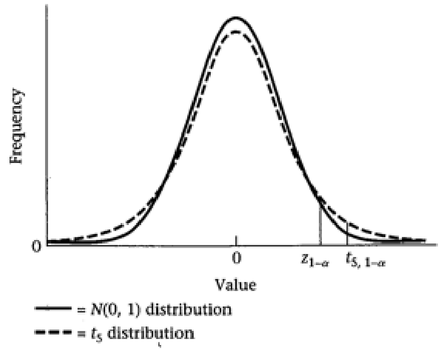
$$P(t_{n-1,\alpha/2} < t < t_{n-1,1-\alpha/2}) = 1 - \alpha$$

The above equation leads to

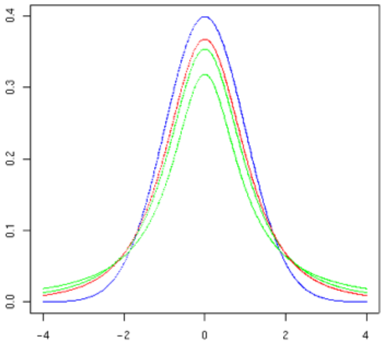
$$P(\bar{X} - t_{n-1,\alpha/2}S/\sqrt{n} < \mu < \bar{X} + t_{n-1,\alpha/2}S/\sqrt{n}) = 1 - \alpha$$

The interval $[\bar{X} - t_{n-1,\alpha/2}S/\sqrt{n}, \bar{X} + t_{n-1,\alpha/2}S/\sqrt{n}]$ is referred to as a $100(1-\alpha)\%$ confidence interval for μ.

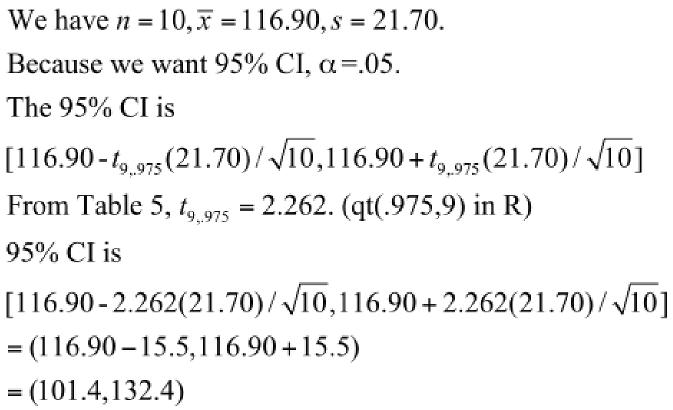
Comparison of Student's *t* distribution with 5 degrees of freedom with an *N*(0, 1) distribution



- T distribution with degree of freedom 1, 2, 3 and Normal (blue).



4. Compute a 95% CI for the mean birthweight based on the sample of size 10 in the previous example.



We have $n = 10, \bar{x} = 116.90, s = 21.70.$

Because we want 95% CI, α =.05.

The 95% CI is

$[116.90 - t_{9,.975}(21.70)/\sqrt{10}, 116.90 + t_{9,.975}(21.70)/\sqrt{10}]$

From Table 5, $t_{9,.975} = 2.262$. (qt(.975,9) in R)

95% CI is

$[116.90 - 2.262(21.70)/\sqrt{10}, 116.90 + 2.262(21.70)/\sqrt{10}]$

$= (116.90 - 15.5, 116.90 + 15.5)$

$= (101.4, 132.4)$

5. Using confidence interval in decision making: Suppose we know the mean cholesterol level in children ages 2-14 is 175 mg/dL. We wish to see if there is a familial aggregation of cholesterol levels. Identify a group of fathers with cholesterol levels $\geq 250 mg/dL$ and measure the cholesterol levels of their 2-14-year-old offspring. Suppose we find the mean cholesterol level in a group of 100 such children is 207.3 mg/dL with standard deviation = 30 mg/dL. Is this value far enough from 175 mg/dL for us

to believe that the underlying mean cholesterol level in the population of all children is different from 175 mg/dL?
Construct a 95% CI for μ on the basis of our sample data. Decision rule: if the interval contains 175 mg/dL, then we cannot say the underlying mean for this group is any different from the mean for all children (175). If the CI does not contain 175, then we would conclude the true underlying mean for this group is different from 175. If the lower bound of the CI is above 175, then there is a demonstrated familial aggregation of cholesterol levels. The CI in this case is given by

$$207.3 \pm t_{99,0.975}(30)/\sqrt{100} = 207.3 \pm 6.0 = (201.3, 213.3)$$

2 Case study

1. Assess whether there is a relationship between bone-mineral density (BMD) and cigarette smoking. 41 twin pairs are selected and each pair has different smoking histories. Matched-pair study - The exposed (heavier-smoking twin) and control (lighter-smoking twin) are matched on other characteristics related to the outcome (BMD). The matching is based on having similar genes so that the effect of genes on outcome can be safely ignored.

2. The difference in bone-mineral density (BMD) was studied as function of the difference in tobacco use. Difference: BMD of heavier-smoking twin - BMD of lighter-smoking twin. Tobacco consumption was expressed in terms of pack-year. One pack-year is defined as 1 pack of cigarettes per day consumed for a year. BMD was assessed separately at three sites - The lumbar spine, the femoral neck, and the femoral shaft.

3. To assess whether there is a relationship between BMD and cigarette smoking. Calculate the difference in BMD between heavier-smoking twin and the lighter-smoking twin for each twin pair. Calculate the average of these differences, which is $-0.036 \pm 0.014 g/cm^2$.

4. The 95% CI for the true mean difference in BMD is

$$-0.036 \pm t_{40,0.975}(s/\sqrt{41}) = -0.036 \pm 2.021(0.014) = (-0.064, -0.008)$$

5. Upper bound is less than 0. The true mean difference is less than 0. The true mean BMD for the heavier-smoking twins is lower than for the lighter-smoking twins.