Correlation & Estimation - Class 7

January 28, 2014 Debdeep Pati

Association between two variables

1. Covariance between two variables X and Y is denoted by Cov(X, Y) and defined by

$$Cov(X,Y) = E(X - E(X))(Y - E(Y))$$

2. Covariance is not convenient for expressing the strength of association between two variables.

3. The correlation coefficient between 2 random variables X and Y is denoted by $Corr(X,Y)$ and is defined by

$$\rho = corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

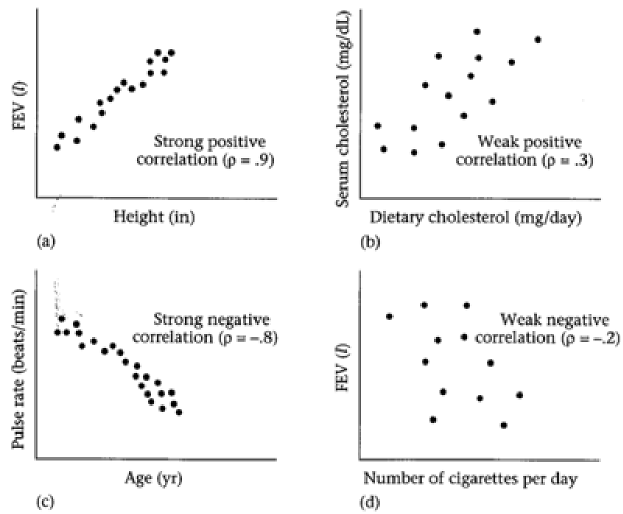
4. ρ is a dimensionless quantity between -1 and 1, for linearly related random variables, 0 implies independence In general, correlation zero does not necessarily imply independence

5. 1 implies nearly perfect positive dependence, -1 implies nearly perfect negative dependence.

Cumulative distribution function

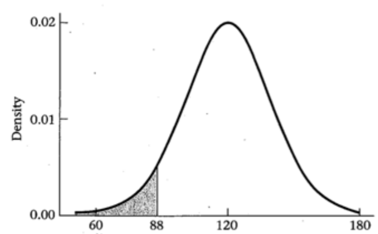
Cumulative distribution function (cdf) for the random variable X evaluated at the point a is defined as the probability that X will take on values $\leq$ a. It is represented by the area under the pdf to the left of a.

Figure 5.16 Interpretation of various degrees of correlation



40

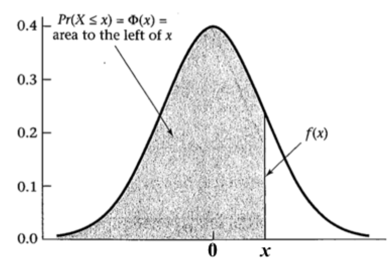
Figure 5.3 Probability-density function for birthweight



43

(a) cdf

Cumulative-distribution function [$\Phi(x)$] for a standard normal distribution



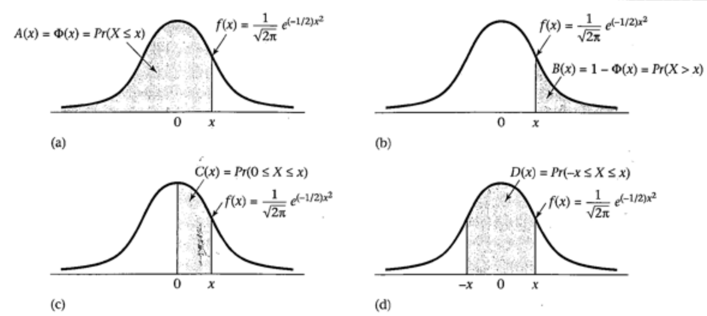
(b) cdf of normal

Normal Table

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

Table 3 The normal distribution



x	A^a	B^b	C^c	D^d	x	A	B	C	D
0.0	.5000	.5000	.0	.0	0.32	.6255	.3745	.1255	.2510
0.01	.5040	.4960	.0040	.0080	0.33	.6293	.3707	.1293	.2586
0.02	.5080	.4920	.0080	.0160	0.34	.6331	.3669	.1331	.2661
0.03	.5120	.4880	.0120	.0239	0.35	.6368	.3632	.1368	.2737
0.04	.5160	.4840	.0160	.0319	0.36	.6406	.3594	.1406	.2812
0.05	.5199	.4801	.0199	.0399	0.37	.6443	.3557	.1443	.2886
0.06	.5239	.4761	.0239	.0478	0.38	.6480	.3520	.1480	.2961
0.07	.5279	.4721	.0279	.0558	0.39	.6517	.3483	.1517	.3035
0.08	.5319	.4681	.0319	.0638	0.40	.6554	.3446	.1554	.3108
0.09	.5359	.4641	.0359	.0717	0.41	.6591	.3409	.1591	.3182
0.10	.5398	.4602	.0398	.0797	0.42	.6628	.3372	.1628	.3255
0.11	.5438	.4562	.0438	.0876	0.43	.6664	.3336	.1664	.3328
0.12	.5478	.4522	.0478	.0955	0.44	.6700	.3300	.1700	.3401

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

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

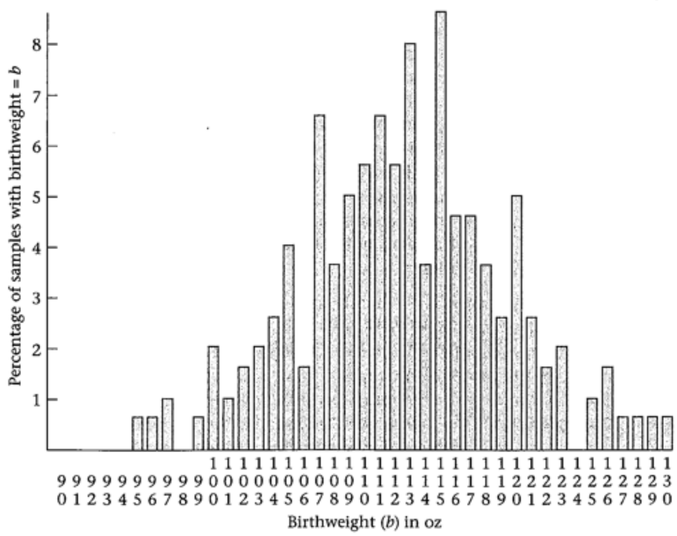
2. Since each X_i's are assumed to be random variables, the quantity $\bar{X}$ is also random.

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.

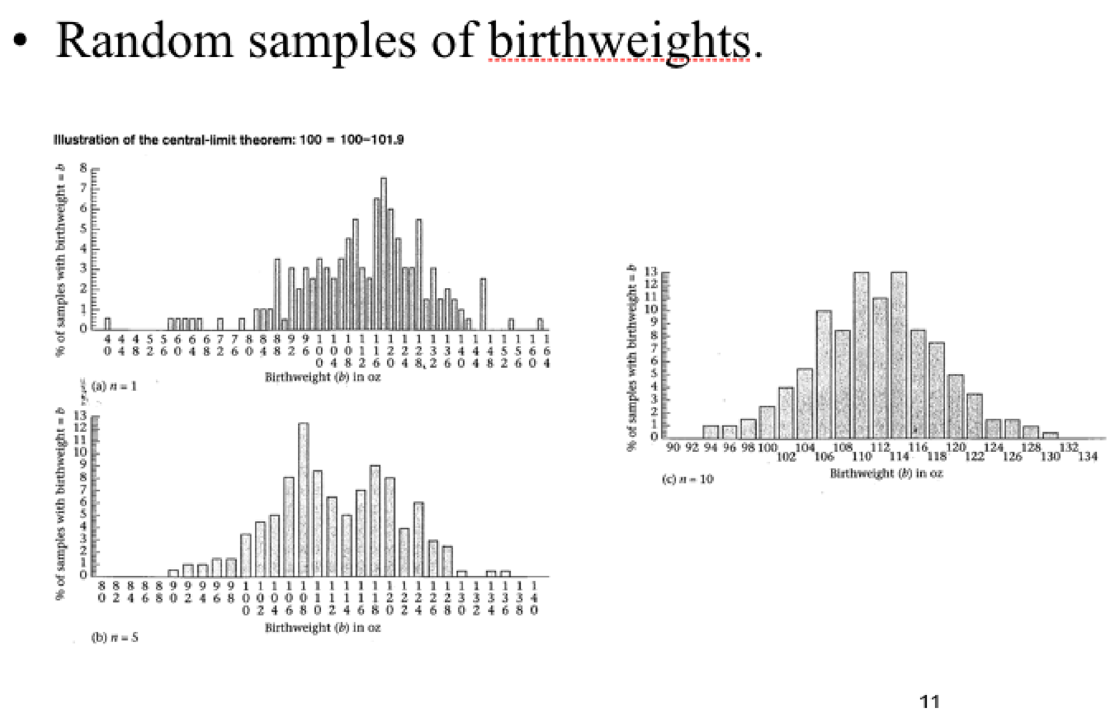
8. Ex. Compute the standard error of the mean for the following sample of birth weights. $97, 125, 62, 120, 132, 135, 118, 137, 126, 118$.

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

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

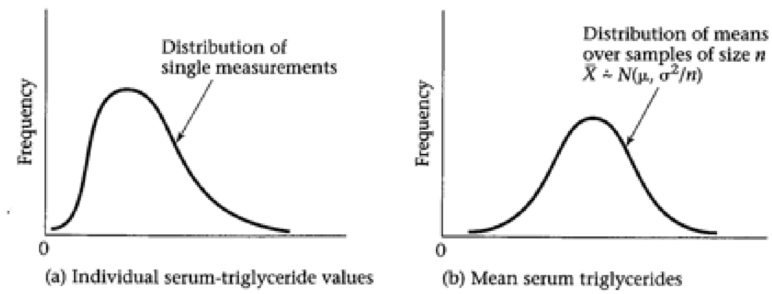
Central Limit Theorem

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.



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

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



12

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 $\sigma/\sqrt{n} = 20.6\sqrt{10} = 6.51$. 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.