Review 2 (on Chapters 3 — 9)

April 13, 2011

1 Chapter 3: Probability

Different ways of defining probability

1. Classical approach 2. frequency approach

3. subjective approach 4. Axioms

Basic tools and technics

- 3 Axioms
- exclusive and complement events
- conditional probability
- independent events
- law of total probability
- Bayes theorem

Example A coin with P(H) $= 0.7$, P(T) $= 0.3$. The coin is tossed repeatedly until Head appears. Find the probability that the first H appears in the 3rd toss.

$$P(TTH) = P(T)P(T)P(H) = 0.3 * 0.3 * 0.7 = 0.0630$$

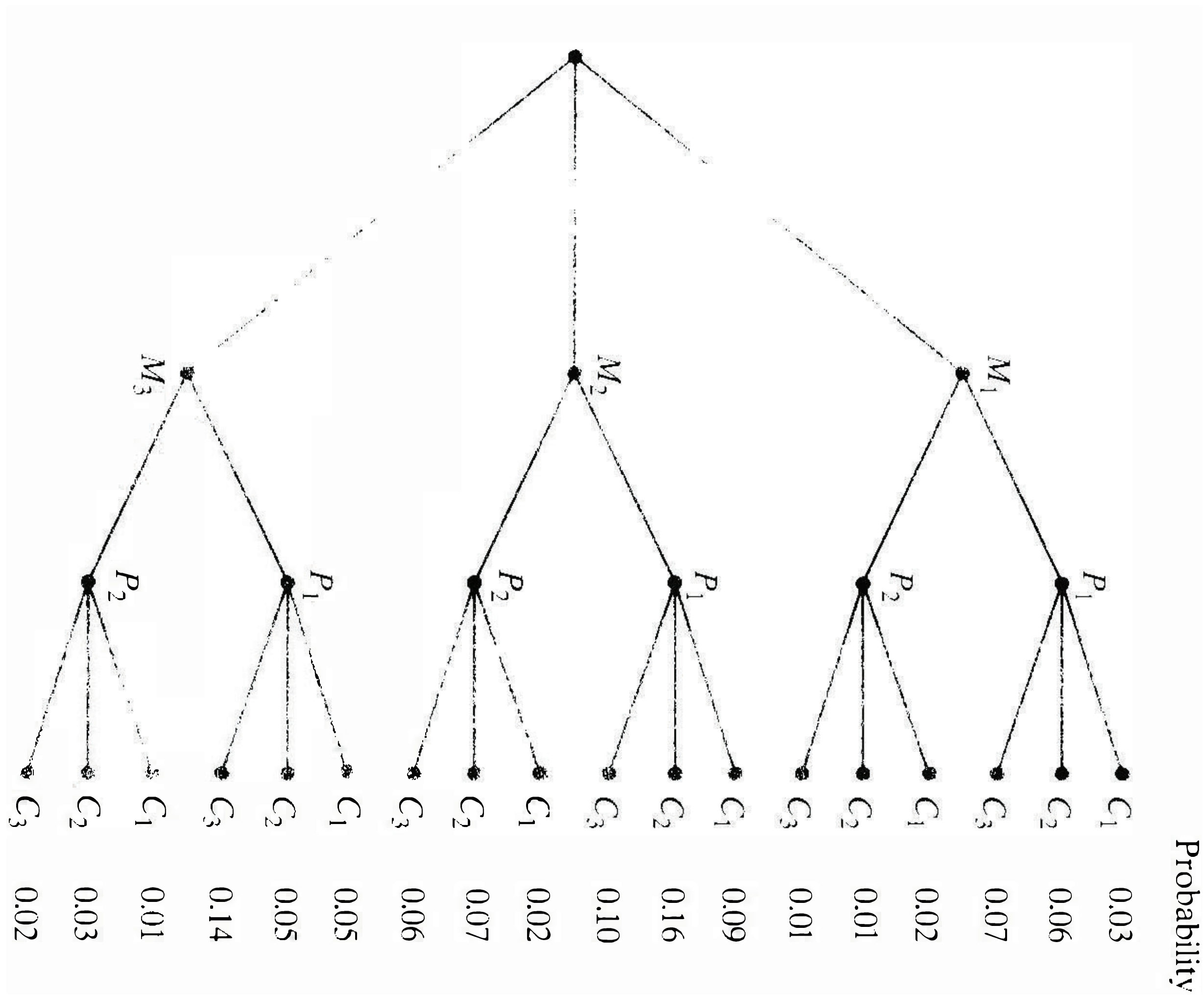
Example A convict is planning to escape prison by leaving through the prison sewer system. There are 4 manholes in the prison, but only one leads to the outside. The prisoner plans to enter each of the manholes at random, never reentering an unsuccessful manhole. What is the probability (1) that the prisoner must try exactly 1 manhole? and (2) the probability that the prisoner can get out before trying 3rd manholes?

Denote by A the even he selects the right hole leading to the outside

(1) P(A) $= 1/4$

(2) $P(A) + P(\bar{A}A) = 1/4 + P(\bar{A})P(A|\bar{A}) = 1/4 + \frac{3}{4} \times \frac{1}{3} = 1/2$

Example



M_3
M_2
M_1
P_2
P_1
C_3 C_2 C_1
Probability
0.02 0.03 0.01 0.14 0.05 0.05 0.06 0.07 0.02 0.10 0.16 0.09 0.01 0.01 0.02 0.07 0.06 0.03

$$P(M_1) = 0.03 + 0.06 + 0.07 + 0.02 + 0.01 + 0.01 = 0.20$$

$$P(P_1) = 0.03 + 0.06 + 0.07 + 0.09 + 0.16 + 0.10 + 0.050.05 + 0.14 = 0.75$$

$$P(C_3) = 0.07 + 0.01 + 0.10 + 0.06 + 0.14 + 0.02 = 0.40$$

$$P(M_1 \cap P_1) = 0.03 + 0.06 + 0.07 = 0.16$$

$$P(M_1 \cap C_3) = 0.07 + 0.01 = 0.08$$

$$P(C_3 \cap P_2) = 0.02 + 0.06 + 0.02 = 0.09$$

$$P(M_1|P_1) = P(M_1 \cap P_1)/P(P_1) = 0.16/0.75$$

$$P(C_3|P_2) = P(C_3 \cap P_2)/P(P_2) = 0.09/(1 - P(P_1)) = 0.09/(1 - 0.75)$$

Example A clinic has 3 doctors Dr A sees 41% of the patients and sent 5% of his patients for blood test. Dr B sees 32% of the patients and requested blood test on 8% of her patients. Dr C sees the rest and sent 6% of his patient for blood test. What is the probability that a blood test request was sent by Dr A?

P(A) = 0.41, P(B) = 0.32, P(C) = 1-0.41-0.32 = 0.27 and $P(T|A) = 0.05, P(T|B) = 0.08, P(T|C) = 0.06$. To find $P(A|T)$

$$\begin{aligned} P(A|T) &= \frac{P(A)P(T|A)}{P(A)P(T|A) + P(B)P(T|B) + P(C)P(T|C)} \\ &= \frac{0.41 * 0.05}{0.41 * 0.05 + 0.32 * 0.08 + 0.27 * 0.06} = 0.3291 \end{aligned}$$

2 Chapters 4-5: Random Variable and Their Distributions

5 important distributions

- Bernoulli distribution/trial

x	0	1
$P(X = x)$	p	1-p

- Binomial distribution $X \sim B(n, p)$
- Poisson distribution $X \sim Poi(\lambda)$
- Normal distribution $X \sim N(\mu, \sigma^2)$
- t-distribution $X \sim t(\nu)$

For the first 4 distributions,

- How to calculation the probability, e.g. $P(X = i)$ or $P(a < X \leq b)$

- what situations a random will follow the distributions?

- expectation and variance of the distribution (or random variables).

For the last 2 distributions (Normal and t-distribution)

- The symmetry of the probability density functions

- how to use the statistical tables?

- what sample statistics follow the distributions?

- For normal distribution, how to Standardize it? (i.e. by subtracting its mean and dividing its standard deviation) $X \sim N(\mu, \sigma^2)$, Then

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

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

1. $P(Z > 2) =??$

2. $P(|Z| < 1.96) =??$

3. $P(Z <??) = 0.05$

4. $P(|Z| >??) = 0.05$

Example If $Z \sim N(1, 2)$

1. $P(Z > 2) = ??$

2. $P(Z < ??) = 0.05$

Example $t \sim t(6)$

1. $P(|t| > ??) = 0.05$ or 0.01

2. $P(t < ??) = 0.05$ or 0.01

3. $P(t > ??) = 0.05$ or 0.01

symbols $z_\alpha, z_{\alpha/2}, -z_\alpha, -z_{\alpha/2}, t_\alpha, t_{\alpha/2}, -t_\alpha, -t_{\alpha/2}$

3 Chapter 6: Sampling distribution, the basics for CI and Tests

For one population,

- If $X_1, ..., X_n$ are independent and each of them follows $N(\mu, \sigma^2)$, then

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- CLT, If $X_1, ..., X_n$ are independent and each of them has mean μ and variance σ^2, then

$$\bar{X} \approx N(\mu, \sigma^2)$$

providing $n > 30$.

- For both cases, denote its standardized version by Z

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \approx N(0,1)$$

- If σ^2 is unknown, we replace σ^2 by S^2 and denote

$$t = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1), \qquad \text{if } n \leq 30$$

or

$$Z = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx N(0,1), \qquad \text{if } n > 30$$

4 Chapters 7-9: CI and Tests for different cases

The parameters we are interested are

$$\mu, \quad p, \quad \sigma^2$$

(but CI and test about σ^2 is not discussed)

All the calculation are based on different formula and distributions.

- **One population: μ and p**

(a) CI for μ

(b) CI for p

(c) test for $H_0 : \mu = \mu_0$

(d) test for $H_0 : p = p_0$

- **Two population: $\mu_1 - \mu_2$ and $p_1 - p_2$**

(i) CI for $\mu_1 - \mu_2$

(ii) CI for $p_1 - p_2$

(iii) test for $H_0 : \mu_1 = \mu_2$ (for independent samples)

(iv) test for $H_0 : p_1 = p_2$

We discuss the cases separately

(a) – distribution (4 cases: n big, n small, σ^2 is know and unknown)

$$\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \approx N(0,1), \quad \frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1), \quad \frac{\bar{X}-\mu}{\sqrt{S^2/n}} \approx N(0,1)$$

– CI of $100(1-\alpha)\%$

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad \bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}, \qquad \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

(b) – distribution (only one case is practical)

$$- - - - - - - - - -, \quad - - - - - - - - - -, \quad \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx N(0,1)$$

– CI of $100(1-\alpha)\%$

$$- - - - - - - - - -, \quad - - - - - - - - - -, \quad \bar{x} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

(c) – distribution under $H_0 : \mu = \mu_0$ (4 cases as above)

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \approx N(0,1), \quad t = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \sim t(n-1), \quad Z = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \approx N(0,1)$$

– regions of rejecting H_0

* $H_1 : \mu \neq \mu_0$

$$|z| > z_{\alpha/2}, \qquad |t| > t_{\alpha/2}, \qquad |z| > z_{\alpha/2}$$

* $H_1 : \mu > \mu_0$

$$z > z_\alpha, \qquad t > t_\alpha, \qquad z > z_\alpha$$

* $H_1 : \mu < \mu_0$

$$z < -z_\alpha, \qquad t < -t_\alpha, \qquad z < -z_\alpha$$

(d) – distribution under $H_0 : p = p_0$ (only one practical case)

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \approx N(0, 1), \quad \text{— — — — — — — — —}, \quad \text{— — — — — — — —}$$

– regions of rejecting H_0

$*$ $H_1 : \mu \neq \mu_0$

$$|z| > z_{\alpha/2}, \quad \text{— — — — — — — — —}, \quad \text{— — — — — — — — —}$$

$*$ $H_1 : \mu > \mu_0$

$$z > z_\alpha, \quad \text{— — — — — — — — —}, \quad \text{— — — — — — — — —}$$

$*$ $H_1 : \mu < \mu_0$

$$z < -z_\alpha, \quad \text{— — — — — — — — —}, \quad \text{— — — — — — — — —}$$

(i) – distribution (2 practical cases)

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_1^2/n_1+S_2^2/n_2}} \approx N(0,1), \qquad \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_p^2/n_1+S_p^2/n_2}} \approx N(0,1)$$

– CI of $100(1-\alpha)\%$

$$\bar{x}-\bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, \qquad \bar{x}-\bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}$$

(ii) – distribution (1 practical case)

$$\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1+\hat{p}_2(1-\hat{p}_2)/n_2}} \approx N(0,1), \qquad - - - - - - - - -$$

– CI of $100(1-\alpha)\%$

$$\hat{p}_1-\hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \qquad - - - - - - - - -$$

(iii) – distribution under $H_0 : \mu_1 = \mu_2$ (when $n_1 > 30, n_2 > 30$ with 2 cases)

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \approx N(0,1), \quad Z = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2/n_1 + S_p^2/n_2}} \approx N(0,1),$$

– regions of rejecting H_0

$* \ H_1 : \mu \neq \mu_0$

$$|z| > z_{\alpha/2}, \qquad |z| > z_{\alpha/2}$$

$* \ H_1 : \mu > \mu_0$

$$z > z_\alpha, \qquad z > z_\alpha$$

$* \ H_1 : \mu < \mu_0$

$$z < -z_\alpha, \qquad z < -z_\alpha$$

(iv) – distribution under $H_0 : p_1 = p_2$ [when $n_1 > 30, n_2 > 30$]

$$- - - - - - - - -, \quad Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2}} \approx N(0,1),$$

– regions of rejecting H_0

$* \; H_1 : \mu \neq \mu_0$

$$- - - - - --, \qquad |z| > z_{\alpha/2}$$

$* \; H_1 : \mu > \mu_0$

$$- - - - - --, \qquad z > z_\alpha$$

$* \; H_1 : \mu < \mu_0$

$$- - - - - --, \qquad z < -z_\alpha$$

(v) For two population but paired observations.

After taking difference, this case is equivalent to the case of (a) and (c)

Example A sample of 150 people was randomly drawn. Each person was identified as a consumer or a non-consumer of high-fiber cereal. For each person the number of calories consumed at lunch was recorded. The data:

	sample from high fiber cereal consumer	sample from non-consumer
sample size	$n_1 = 43$	$n_2 = 107$
sanple mean	$\bar{x} = 604.02$	$\bar{y} = 633.23$
sample variance	$s_1^2 = 50680$	$s_2^2 = 51040$

- find the confidence intervals for the difference of calories for the two groups with 95% confidence

- test whether people who eat high-fiber cereal take less calories at significance level 1%.

Solution:

Since $s_1/s_2 = 0.9965$ is in between (0.5, 2), the two population have the same variance. The pooled sample variance is

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{42 * 50680 + 106 * 51040}{43+107-2} = 50938$$

- When $\alpha = 0.05$ we have $z_{\alpha/2} = 1.96$. The CI with 95% confidence

$$\begin{aligned} \bar{x} - \bar{y} \pm z_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &= 604.02 - 633.23 \pm 1.96\sqrt{50938}\sqrt{\frac{1}{43} + \frac{1}{107}} \\ &= [-109.0822, 50.6622] \end{aligned}$$

- $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 < \mu_2$ with $\alpha = 0.01$. Consider

$$Z = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The rejection region is

$$Z < -z_{0.01} = -2.33$$

The observed value of Z is

$$z = \frac{604.02 - 633.23}{\sqrt{50938}\sqrt{\frac{1}{43} + \frac{1}{107}}} = -0.7168$$

Since $z > -2.33$, we dont reject H_0. People who eat high-fiber cereal does not take significant less calories than those who does not eat high-fiber cereal.