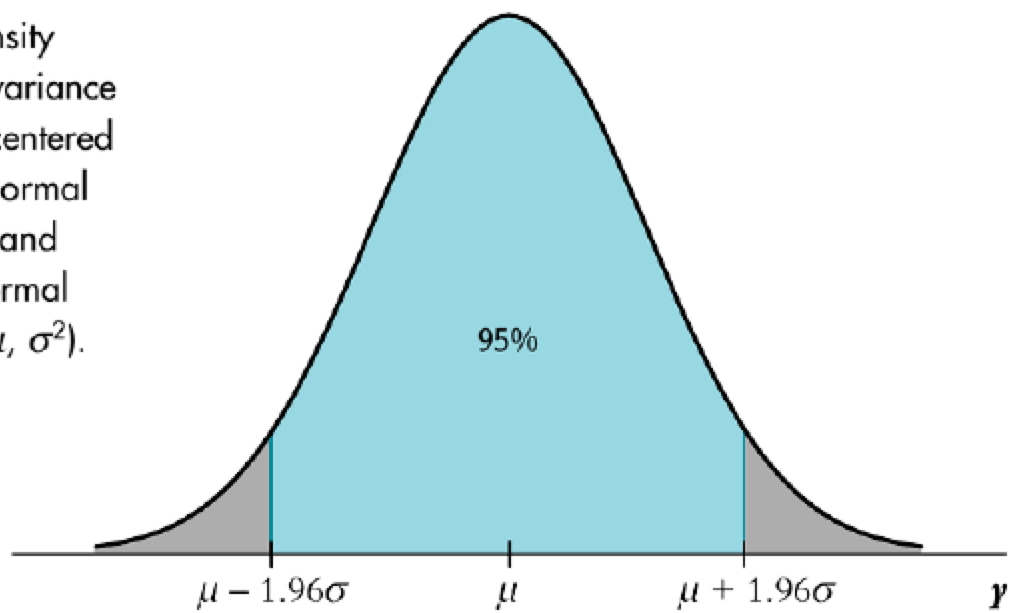
Review of Probability Theory-Part B

4) Distribution

- Normal Distribution $X \sim N(\mu, \sigma^2)$

FIGURE 2.3 The Normal Probability Density

The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.



- $$\Pr(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$$

4) Distribution

- In general, we standardize the Normal R.V. by subtracting μ and dividing by σ.

$$\Pr(c < X < d)$$

$$= \Pr\left(\frac{c-\mu}{\sigma} < \underbrace{\frac{X-\mu}{\sigma}}_{Z} < \frac{d-\mu}{\sigma} \right) = \Pr\left(\frac{c-\mu}{\sigma} < Z < \frac{d-\mu}{\sigma} \right)$$

$$\text{where } Z \sim N(0,1)$$

4) Distribution

- Example: Income in the US is the normally distributed with mu=50,000 and sigma=20,000. What is the probability a person in the US has income over 75,000.

$$\Pr(X > 75{,}000) = \Pr\left(\frac{X - 50{,}000}{20{,}000} > \frac{75{,}000 - 50{,}000}{20{,}000}\right)$$
$$= \Pr(Z > 1.25) = 10.56\%$$

- What is the probability that a person at random has income between 35,000 and 65,000?

$$\Pr(35{,}000 < X < 65{,}000)$$
$$= \Pr\left(\frac{35{,}000 - 50{,}000}{20{,}000} < \frac{X - 50{,}000}{20{,}000} < \frac{65{,}000 - 50{,}000}{20{,}000}\right)$$
$$= \Pr(-0.75 < Z < 0.75) = 1 - \Pr(Z < -0.75) - \Pr(Z > 0.75)$$
$$= 1 - 2 \cdot \Pr(Z < -0.75) = 1 - 2 \cdot 0.2266 = 0.5468 = 54.68\%$$

4) Distribution

- Bi-variate Normal: this describes the probability distribution of X and Y are jointly as Normal

- The linear combination aX+bY is distributed as:

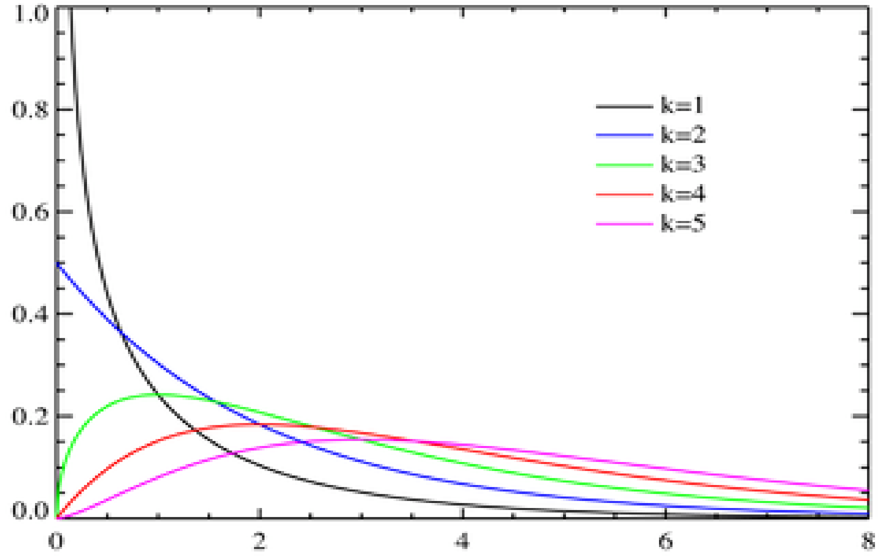
$$N\left(a\mu_X + b\mu_{Y,}\;\; a^2 \cdot \sigma_X^2 + b^2 \cdot \sigma_Y^2 + 2ab \cdot \sigma_{XY}\right)$$

- If X and Y are independent, then aX+bY is distributed as:

$$N\left(a\mu_X + b\mu_{Y,}\;\; a^2 \cdot \sigma_X^2 + b^2 \cdot \sigma_Y^2\right)$$

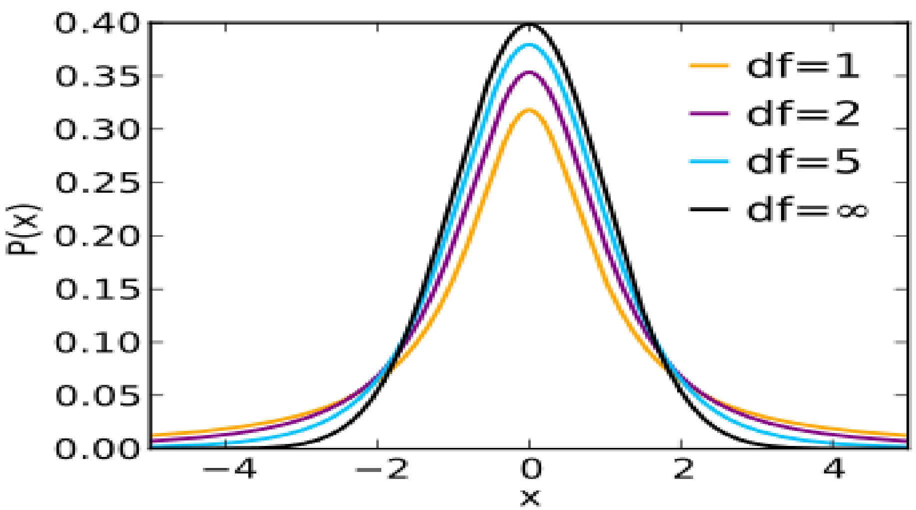
4) Distribution

- Chi-squared distribution χ^2
- The distribution of the sum of k squared independent standard normal RV is called χ_k^2 distribution , and k is called the degree of freedom.
- Suppose that we have 4 (i.e., k=4) standard normal variable that are independent: Z_1, Z_2, Z_3, Z_4. Let $W = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2$. Then the distribution of W is χ_4^2
- $\Pr(W > 9.49) = 0.05$, see table 3 on page 758
- PDF:



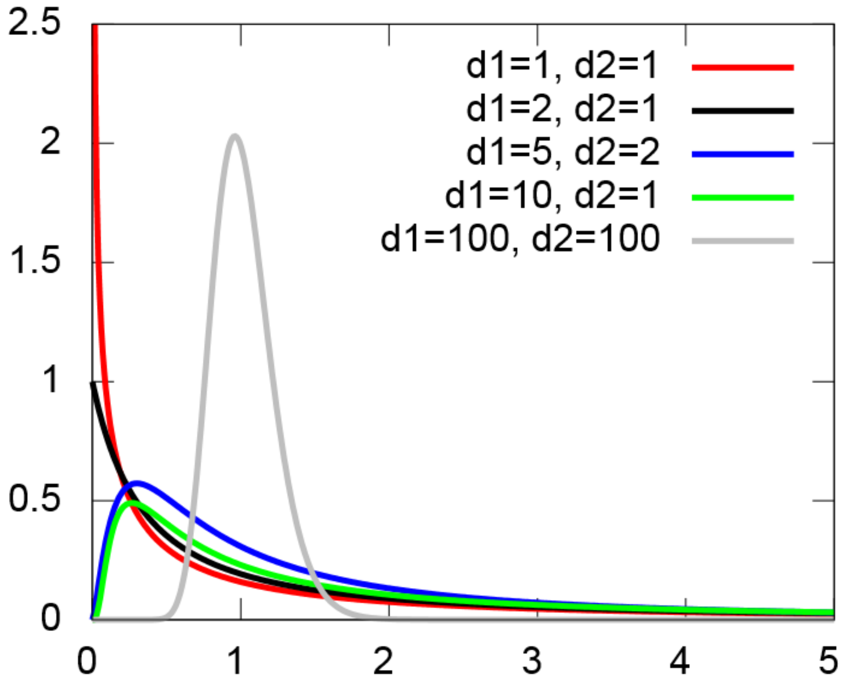
4) Distribution

- Student t distribution
- Let Z be a standard normal RV and Wm a Chi-squared RV with degree of freedom m. Suppose Z and Wm are independent.
- Then $R_m = \dfrac{Z}{\sqrt{\frac{W_m}{m}}} \sim t$ distribution with m degree of freedom
-



4) Distribution

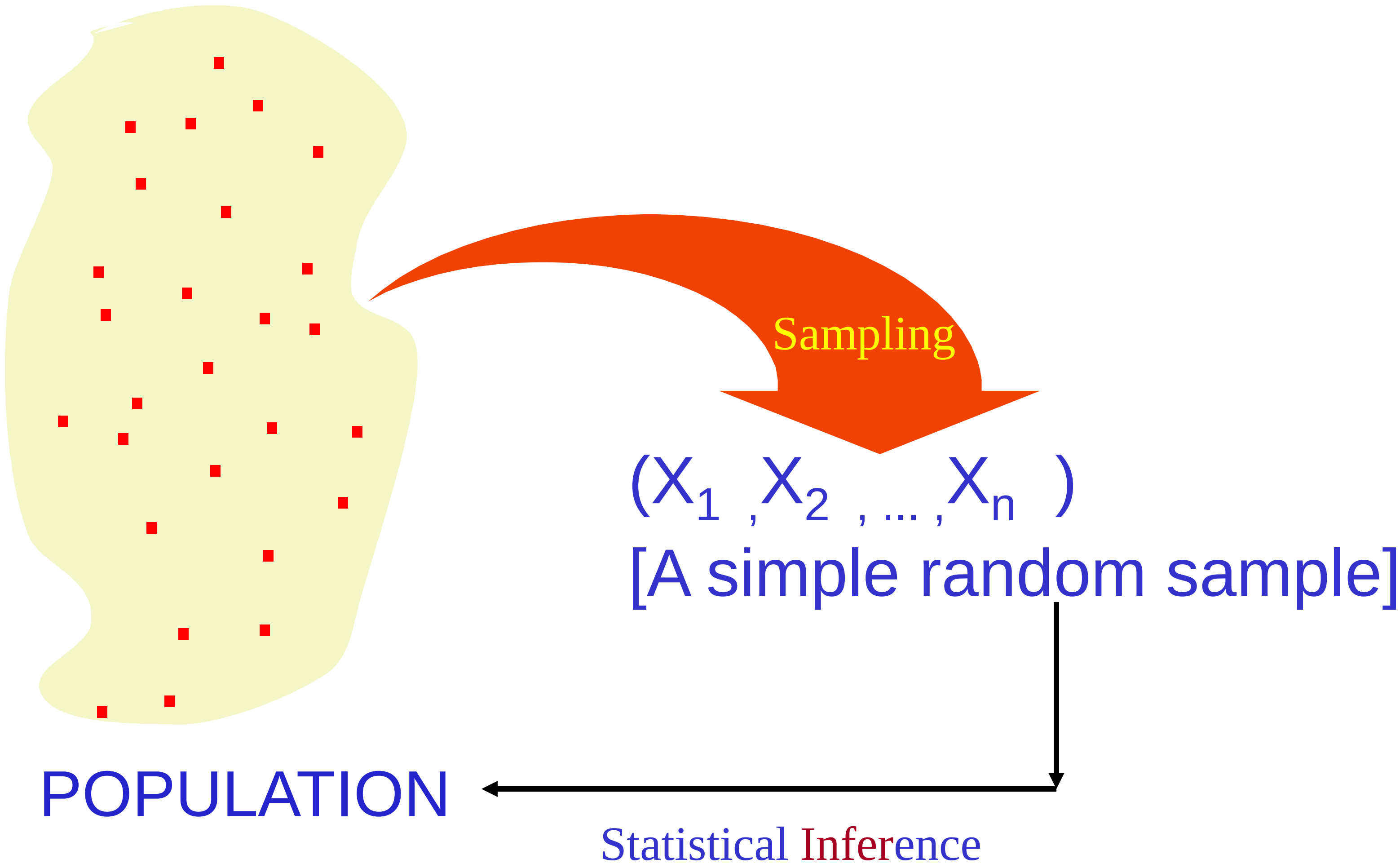
- F distribution
- Let W be a chi-squared random variable with m degree of freedom.
- Let V be a chi-squared random variable with n degree of freedom.
- $\dfrac{W/m}{V/n} \sim F_{m,n}$ (F distribution with m and n degree)

  As $n \to \infty$, $W/m \sim F_{m,\infty}$
- 

5) Sampling Distribution

Sampling is the process of taking a smaller group of subjects from a larger population.

5) Sampling Distribution



5) Sampling Distribution

- The random sample or the data $\{X_1, X_2, X_3 ..., X_n\}$ are random variables.
- We don’t know the exact outcome beforehand.

- The statistic calculated from a randomly chosen sample is an example of a random variable.
- A statistic from different random samples will take different values.

5) Sampling Distribution

- Populations Have Parameters, Samples Have Estimators

Population	Sample
Parameters	Estimators
e.g. Pop. mean *Pop. Variance*	*Sample mean* *Sample variance*

5) Sampling Distribution

- Suppose $Y_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$
- What is the distribution of $\bar{Y} = \frac{\sum Y_i}{n}$?
- Here we focus on the case where a sample is draw at random (random sampling) from a population
- For a sample size n=20, draw at random $Y_1, Y_2, \ldots, Y_{20}$ each Y_i is independent from the other since random sampling
- If they are draw from the same underling distribution (e.g., Normal), then they are identically distributed
- iid: independently and identically distributed.

$$Y_1, Y_2, Y_3, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

5) Sampling Distribution

- Average $\bar{Y} = \frac{\sum Y_i}{n}$
- For different random sample drawn from the population implies that $\bar{Y}$ is a random variable
- $\bar{Y}$ changes from sample to sample
- So we are interested in that variability called the sampling distribution
- Suppose $Y_1, Y_2, Y_3, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$
- Then

$$\mu_{\bar{Y}} = E(\bar{Y}) = E\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n}E[Y_1 + Y_2 + \ldots + Y_n]$$

$$= \frac{1}{n}[E(Y_1) + E(Y_2) + \ldots + E(Y_n)] = \frac{1}{n}[\mu + \mu + \ldots + \mu] = \mu$$

$$\sigma^2_{\bar{Y}} = var(\bar{Y}) = var\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n^2}var[Y_1 + Y_2 + \ldots + Y_n]$$

$$= \frac{1}{n^2}[var(Y_1) + var(Y_2) + \ldots + var(Y_n)] = \frac{1}{n^2}[\sigma^2 + \sigma^2 + \ldots + \sigma^2] = \frac{\sigma^2}{n}$$

5) Sampling Distribution

- Example: Incomes normally distributed mu=50,000 and sigma=20,000. 100 people sampled at random from the distribution: $\bar{Y} \sim N\left(50,000, \frac{20,000^2}{100}\right)$
- What is the probability the average income is greater than 55,000?

$$\Pr(\bar{Y} > 55,000) = \Pr\left(\frac{\bar{Y} - 50,000}{\frac{20,000}{\sqrt{100}}} > \frac{55,000 - 50,000}{\frac{20,000}{\sqrt{100}}}\right)$$

$$= \Pr(Z > 2.5) = 0.0062 = 0.62\%$$

- What is the probability the average income is between 45,000 and 52,000.

$$\Pr(45,000 < \bar{Y} < 52,000) = \Pr\left(\frac{45,000 - 50,000}{\frac{20,000}{\sqrt{100}}} < \frac{\bar{Y} - 50,000}{\frac{20,000}{\sqrt{100}}} < \frac{52,000 - 50,000}{\frac{20,000}{\sqrt{100}}}\right)$$

$$= \Pr(-2.5 < Z < 1) = 1 - [\Pr(Z < -2.5) + \Pr(Z > 1)] = 0.8351$$

6) Large sample distribution

- We know that $Y_i \sim N(\mu, \sigma^2),\ i = 1, \ldots, n$, then $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- But if $Y_i \sim$ some complicated distribution, $i = 1, \ldots, n$ then $\bar{Y} \sim\ ?$ Potentially complicated!
- But

$$\text{As } n \to \infty,\ \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

- This is called central limit theorem (CLT).

6) Large sample distribution

- Law of large number
- $Y_1, Y_2, Y_3, \ldots, Y_n$ drawn iid from distribution with mean mu.
- As n goes to infinite, $\bar{Y}$ gets closer and closer to mu.
- Convergence in probability or "consistency"

$$\text{For any } c > 0,\ \Pr(\mu - c < \bar{Y} < \mu + c) \to 1, \text{ as } n \to \infty$$

- Stated as $\bar{Y} \xrightarrow{p} \mu$. This is saying

$$\Pr(\bar{Y} \approx \mu) = 1, \text{ as } n \to \infty$$

$$\Pr(|\bar{Y} - \mu| < c) = 1, \text{ as } n \to \infty$$

$$\Pr(|\bar{Y} - \mu| > c) = 0, \text{ as } n \to \infty$$

6) Large sample distribution

- Law of large number
- Example, $Y_1, Y_2, Y_3, \ldots, Y_n \overset{iid}{\sim} N(\mu = 12, \sigma^2 = 25)$

$$n = 2,\ \bar{Y} = 7.80$$

$$n = 10,\ \bar{Y} = 11.67$$

$$n = 100,\ \bar{Y} = 11.84$$

$$n = 1000,\ \bar{Y} = 11.97$$

As $n \to \infty,\ \bar{Y} \to \mu$

- An interactive example:
- http://digitalfirst.bfwpub.com/stats_applet/generic_stats_applet_11_largenums.htmll

6) Large sample distribution

- Central limit theorem
- $Y_1, Y_2, Y_3, \ldots, Y_n$: iid, then

$$\text{As } n \to \infty,\ \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \overset{A}{\sim} N(0,1) \text{ or}$$

$$\text{As } n \to \infty,\ \bar{Y} \overset{A}{\sim} N\left(u, \frac{\sigma^2}{n}\right)$$

- Very powerful results. It says that even if we don't know the distribution of $Y_1, Y_2, Y_3, \ldots, Y_n$, we know $\bar{Y}$ will be normally distributed if n is sufficiently large.
- The bigger n becomes, the closer $\bar{Y}$'s distribution become normal.
- An interactive example:

http://www.mathsisfun.com/data/quincunx.html