## Correlation and Independence

A book store with a large collection of books, roughly has the following distribution

| Page | Price | proportion |
|---|---|---|
| 100 | 200 | .1 |
| 100 | 1000 | .4 |
| 200 | 1000 | .2 |
| 200 | 200 | .1 |
| 300 | 1000 | .05 |
| 300 | 200 | .05 |
| 400 | 1000 | .03 |
| 400 | 200 | .07 |

$X$ – Number of pages
$Y$ – Price
Joint distribution of $(X, Y)$.

## Correlation and Independence

| Y | 200 | 1000 | |
|---|---|---|---|
| X | | | |
| 100 | .1 | .4 | 0.5 |
| 200 | .1 | .2 | 0.3 |
| 300 | .05 | .05 | 0.1 |
| 400 | .03 | .07 | 0.1 |

$P(Y = 1000|X = 200) = \frac{.2}{.3}$
Conditional distribution of $Y$ given $X$

(Marginal) distn of $X$

| X | prob |
|---|---|
| 100 | 0.5 |
| 200 | 0.3 |
| 300 | 0.1 |
| 400 | 0.1 |

## Independence

$X$, $Y$ are said to be independent if for all $x, y$,

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Another Example

- From a population draw two samples, one $(X_1)$, and then another $(X_2)$
  without replacement – Dependent
- with replacement –Independent

## Independence

- These notions can be extended to continuous random variables.
- Consequences of independence of $X$, $Y$
  - $E(XY) = \mu_X E(Y)$
  - for any $n, m$, $E(X^n Y^m) = E(X^n)E(Y^m)$

# Covariance and Correlation

- $E(XY) - \mu_X E(Y)$ – Covariance between $X$ and $Y$.
- Independence $\Longrightarrow$ Cov(X,Y) =0
- Covariance depends on scale
- Correlation Coefficient

$$\rho(X, Y) = \frac{cov(X, Y)}{s.d(X)s.d(Y)}$$

- $\mathrm{Cov}(aX + b, cY + d) =$ a c $\mathrm{Cov}(X, Y)$. Covariance depends on the scale
- The correlation between $X, Y$ is

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{s.d(X)s.d(Y)}$$

- $\rho(X, Y)$ is scale free, lies between -1 and 1
- $\rho$ is a measure of the linear relationship between $X$ and $Y$. $\rho(X, Y) = 1$implies $Y = aX + b$

# Caution!!

- If $X$ and $Y$ are independent then $\rho(X, Y) = 0$.
- converse is not true. Check the case where $X$ takes values $-1, 0, 1$ with equal probability and $Y = X^2$
- Needs care in interpreting when $\rho(X, Y) \neq 0$. Presence of latent variables.

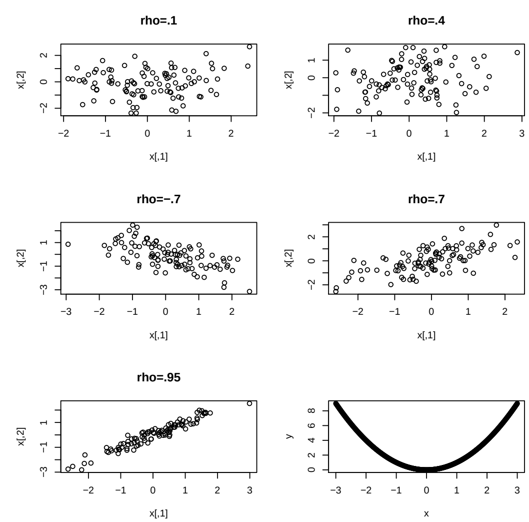



Figure: Correlations

## Sampling distributions

- Suppose we have a large population with average $\mu$ and s.d $\sigma$
- we want to draw $n$ samples from the population and compute their average and s.d
- How is the sample average related to the population average $\mu$

## Sampling distributions

- Before taking the sample, the $n$, sample values $X_1, X_2, \dots, X_n$ that we would obtain are random with distribution governed by the population distribution
- the sample average $\bar{X}$ is also, hence, random
- Since the population is large we may assume that the samples are independent
  - $$E(\bar{X}) = \mu$$
  - $$s.d(\bar{X}) = \frac{\sigma}{n}$$

## Sampling distributions

- Let us interpret the expressions

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{n}$$

- By chebyshev, each of the observations will be within $3\sigma$ of $\mu$ with 90% probability
- The sample average will be within $3\frac{\sigma}{n}$ of $\mu$ with 90% probability.
  In $n = 100$, $\bar{X}$ will be 10 times closer to the population mean with 90% probability.
- Switch the argument. If we do not know $\mu$, we can say that " the population average will be within $3\frac{\sigma}{n}$ of $\bar{X}$ with 90% probability
  This is related to confidence intervals. Shall return later

## A common statistical model

- Think of a Large population – say the people of Kerala.
- We are interested in the aveage income of this population
- Suppose we decide to pick 100 individuals at random from this population and record their incomes
- $X_1$ – income of the first sample is a random variable with distribution given by the income distribution in the population. Similarly $X_2, \dots, X_{100}$ are income of the 100 samples
- Since the population is large, we may assume that $X_1, X_2, \dots, X_{100}$ are independent. Further, they are all samples from the same population, so have the same distribution, in particular the same mean and same s.d

- Applying chebyshev's inequality
  - each sample will have 90% probability of being within 3 s.d of the mean
  - What about $\bar{X}$ ?
  - $\bar{X}$ will be within 3 $\sigma_{\bar{X}}$, 0.3 s.d of the mean with 90% probability

## Earlier slide

- If repeated random samples were drawn from the population with population mean $\mu$ and population s.d. $\sigma$
- average of the data will be approx $\mu$
- s.d. of data will be approx $\sigma$
- in our case
  - if repeated observations were made of $\bar{X}$
  - average of these sample averages will be approx $\mu$
  - s.d. of these sample averages will be approx $s.d(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

## Interpretation

Only proportions matter

Population Distribution

| | |
|---|---|
| 1 | $\times$ 2000 |
| 2 | $\times$ 3000 |
| 3 | $\times$ 4000 |
| 4 | $\times$ 1000 |
| 5 | $\times$ 5000 |
| 6 | $\times$ 5000 |

$\mu = 3.95$
$\sigma$= 1.76

## simulation example

- draw a sample of size 9 from a population with mean , compute the average
- repeat the above say 1000 times. This gives 1000, sample averages
- the average of these “sample averages ” close to 3.95
- the s.d of these “sample averages ” close to 1.76/3 = 0.59

simulation

## Distn of $\bar{X}$ in normal populations

- We have a normal population with mean $\mu$ and s.d $\sigma$
- $\bar{X}$ is the average of $n$ samples from the population
- we have seen
$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{n}$$
- does normality give us anything more?
- $\bar{X}$ is normal with mean $\mu$ and s.d = $\sigma_{\bar{X}} == \frac{\sigma}{\sqrt{n}}$.

## modified slide

- if repeated observations were made of $\bar{X}$
- average of these sample averages will be approx $\mu$
- s.d. of these sample averages will be approx $s.d(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- the histogram of the sample averages will look like a normal with mean $\mu$ and s.d $\frac{\sigma}{\sqrt{n}}$

simulation, normalxbardistn

# A common statistical model-modified

- We are interested in the aveage income of a large population
- Suppose we decide to pick 9 individuals at random from this population and record their incomes
- $X_1$ – income of the first sample is a random variable with distribution given by the income distribution in the population. Similarly $X_2, \ldots, X_9$ are income of the 9 samples

- Applying chebyshev's inequality
  - each sample will have 90% probability of being within 3 s.d of the mean
  - $\bar{X}$ will be within 3 $\sigma_{\bar{X}}$, 0.3 s.d of the mean with 90% probability
- if the population is normal
- each sample will have 95% probability of being within 1.96 s.d of the mean
- we see that $\bar{X}$ will be within 1.96 $\sigma_{\bar{X}}$, $\frac{1.96}{3}$ s.d of the mean with 95% probability

# Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean $\mu$ and finite s.d $\sigma$. Let $\bar{X}_n = \frac{X_1+X_2+\ldots X_n}{n}$. Then for all $t$,

$$P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq t\right) \to \Phi(t)$$

In words, for large $n$, $\bar{X}_n$ is approximately distributed as $N(\mu, \frac{\sigma}{n})$

simulation clt small sample

simulation clt

## CLT and sample proportion

- A population has units that are in one of two categories S, F
- p is the proportion of S
- A sample of size n is drawn
- $X$ number of of S in the sample

$$\hat{p} = \frac{X}{n} \text{ sample proportion}$$

- $\hat{p}$ is approx. $N(p, \sqrt{\frac{p(1-p)}{n}})$

## CLT and sample proportion

- Let $X_1 = 1$ if the first sample is S, and 0 if F
- Let $X_2 = 1$ if the first sample is S, and 0 if F
- ......
- $X = X_1 + X_2 + \ldots + X_n$
- $\hat{p} = \frac{X_1+X_2+\ldots+X_n}{n}$
- Use CLT. Note $E(X_1) = p, \qquad s.d(X_1) = \sqrt{p(1-1p)}$

## CONFIDENCE INTERVALS

- confidence interval for the mean of a normal population
- $\sigma$ known Toy model, but illustrative
- $\sigma$ unknown
- Large sample confidence interval for proportion

## CONFIDENCE INTERVALS

- We have a population that can be modelled as Normal
- The population mean $\mu$ is not known
- The population s.d $\sigma$ is known

## CONFIDENCE INTERVALS

- Using a sample (of size ‘n‘) propose a range of values for $\mu$
- State a measure of confidence of the proposed interval
- A 95% confidence level for $\mu$ means
- We want ‘B‘ such that

$$P(\bar{X} - B < \mu < \bar{X} - B) = .95$$

- how do we find ‘B‘?

## Confidence intervals: known $\sigma$

Since

- $$\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$
- $$P(-1.96 < \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} < 1.96) = .95$$
- A bit of algebra gives

$$P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = .95$$

- If we set $B = 1.96\frac{\sigma}{\sqrt{n}}$, then

$$P(\mu \in \bar{X} \pm B) = .95$$

- $\bar{X} \pm B$ is called 95 % confidence interval for the mean

## Confidence intervals: Interpretation

- Suppose repeated samples are to be drawn.
- In each case we claim that $\bar{X} \pm B$ contains the population mean
- In 95% of the cases we would be right

## Confidence intervals

- In general formulation, we model the population to have a distribution depending on some unknown parameters for normal $\mu, \sigma$
- We want to get a confidence interval for a parameter $\theta$ for normal $\mu, \sigma$
- the general expression is $\hat{\theta} + k_c(s.d(\hat{\theta}))$ , where
  - $\hat{\theta}$ is an estimate of $\theta$
  - $k_c$ is a factor determined by the confidence level and the distribuiton of $\frac{\hat{\theta}-\theta}{s.d(\hat{\theta})}$

## Confidence intervals

- the general expression is $\hat{\theta} + k_c(s.d(\hat{\theta}))$ , where
  - $\hat{\theta}$ is an estimate of $\theta$
  - $k_c$ is a factor determined by the confidence level and the distribuiton of $\frac{\hat{\theta}-\theta}{s.d(\hat{\theta})}$
- In general $s.d(\hat{\theta})$ would involve population parameters and one would substitute it an estimate of $s.d(\hat{\theta})$. This is called STANDARD ERROR
- $\hat{\theta} + k_c(S.E(\hat{\theta}))$

## Confidence intervals

- If both $\mu$ and $\sigma$ of a normal population is not known.
- $s.d(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. and S.E.$(\bar{X}) = \frac{s}{\sqrt{n}}$
- where $s$ is the standard deviation of the sample
- $k_c$ is computed using a $t$ distribution

## Confidence intervals

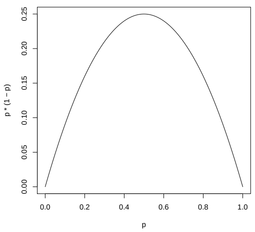
- In a population with objects of two types (S, F). Want to estimate the proportion $p$ of S in the population
- $\hat{p} = \dfrac{\text{number of S in the sample}}{n}$ is an estimate of $p$
- Standard error of $\hat{p} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- for large $n$, $k_c$ is calculated using normal tables.

## Other Issues

- For the normal, suppose we fix B and confidence level 95%. How large a sample do we need to ensure that $\bar{X} \pm B$ has 95% confidence level
- Look at $B = 1.96\frac{\sigma}{\sqrt{n}}$. solve for $n$
- $n = [\frac{1.96\sigma}{B}]^2$

## Other Issues

- In a population with objects of two types (S, F). Want to estimate the proportion $p$ of S in the population
- Suppose we fix B and confidence level 95%. How large a sample do we need to ensure that $\bar{p} \pm B$ has 95% confidence level
- $n = [\frac{1.96\sqrt{\hat{p}}(1-\hat{p})}{B}]^2$




- the S.E. attains maximum at .5. So conservative value for $n$ is $[\frac{1.96\times0.5}{B}]^2$