Descriptive Statistics

Nguyen The Hien, DAH - Vietnam

CONTENT

I. Statistical Method

1. ***Descriptive statistics***
2. *Inferential statistics*

II. Probability.

1. *Definition of probability*
2. *Normal distribution*
3. *Binomial distribution*

I. Statistical Method

1. Descriptive statistics

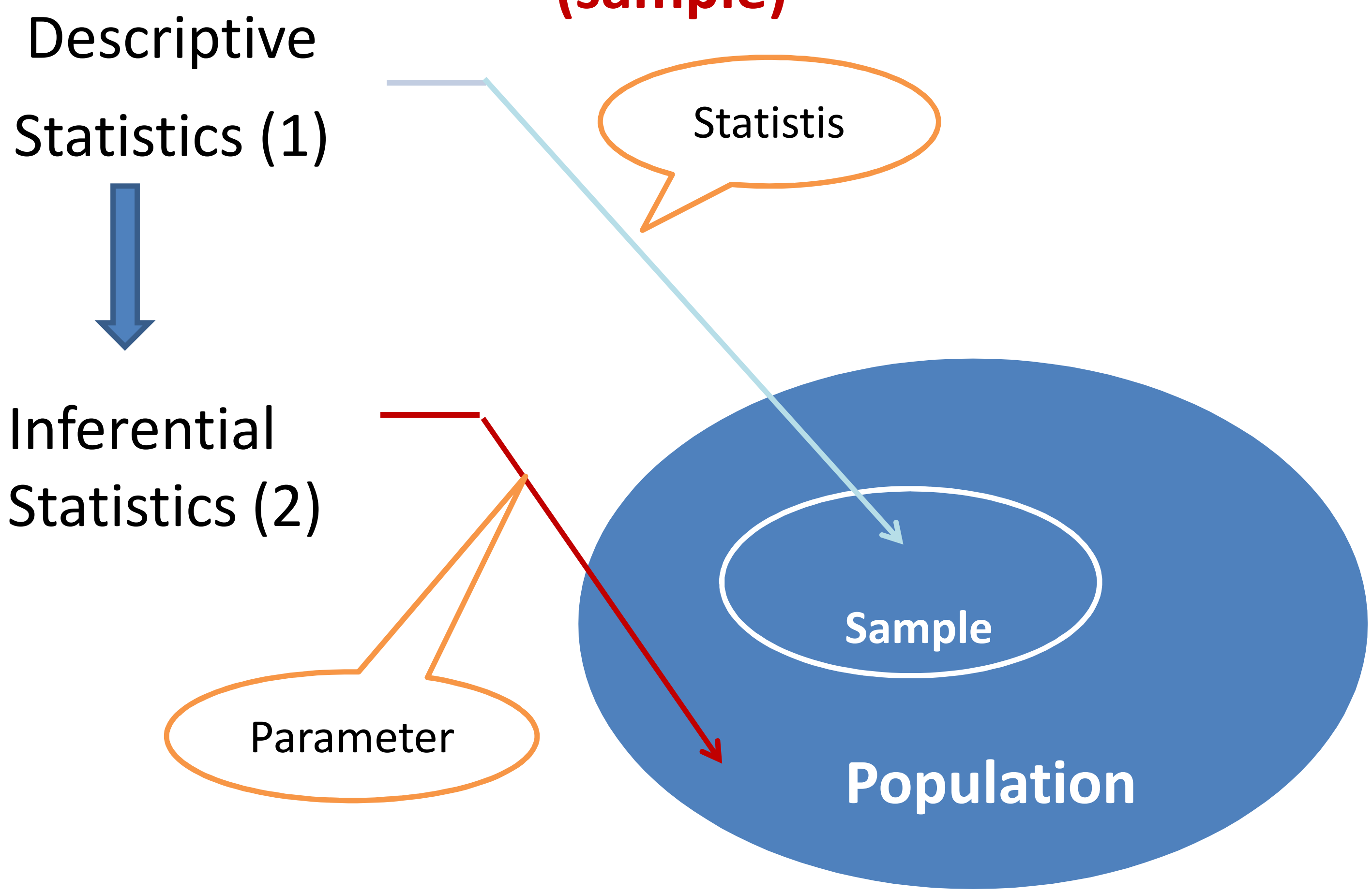
- A method of collecting, organizing, displaying and drawing conclusion of data

2. Inferential statistics

- Use the result of research from sample statistics to population parameter

- Type of inferential statistics methods:
  + Parameter estimation
  + Hypothesis testing

(1) Give a clearer picture of what we want to describe in a group (sample)



(2) Give answer in the world of uncertainty (population)

Summary: Descriptive Statistics

Type of variable	Measurement scale	Statistics	Presentation
Numerical discrete	- Interval - Ratio	Mean, or Median, or Mode (rare), Range, SD, IQR	Histogram scatter plot line graph
Numerical continuous	- Interval - Ratio	Mean, or Median, or Mode (rare), Range, SD, IQR	Histogram scatter plot line graph
Categorical dichotomous	- Nominal - Ordinal	Proportion (%)	Table, Pie, Bar chart
Categorical polychotomous	- Nominal - Ordinal	Proportion (%)	Table, Pie, Bar chart

Notation

	Statistics	Parameter
Mean	$\bar{X}$	μ
Variance	S^2	б^{2}
Standard deviation	S	б
Standard error	$S_{\bar{X}}$	б$_{\bar{X}}$
Proportion	P	π
Correlation coefficient	r	ρ

II. Probability

1. Definition and characteristic

Definition: The chance, or likelihood that an event occurs.

Characteristics:

- Value of probabilities to happen an event will be around 0 – 1
- The total of Pr of all events are 1.0

Ex.

- $Pr\ (x = n) = X_1$ ------> probability of an value/point
- $Pr\ (x < n_a)$ = Sum of $Pr\ (x = n_i)$ *with* $i = 0, 1, \ldots\ n_a$
- $Pr\ (n_a < x < n_b)$ = Sum of $Pr\ (x = n_i)$ *with* $i = a, a+1, \ldots, b$

----> *pace/interval*

2. Probability distribution

- Continuous probability distribution
  - Normal distribution
- Discrete probability distribution
  - Binomial distribution
  - Poisson distribution

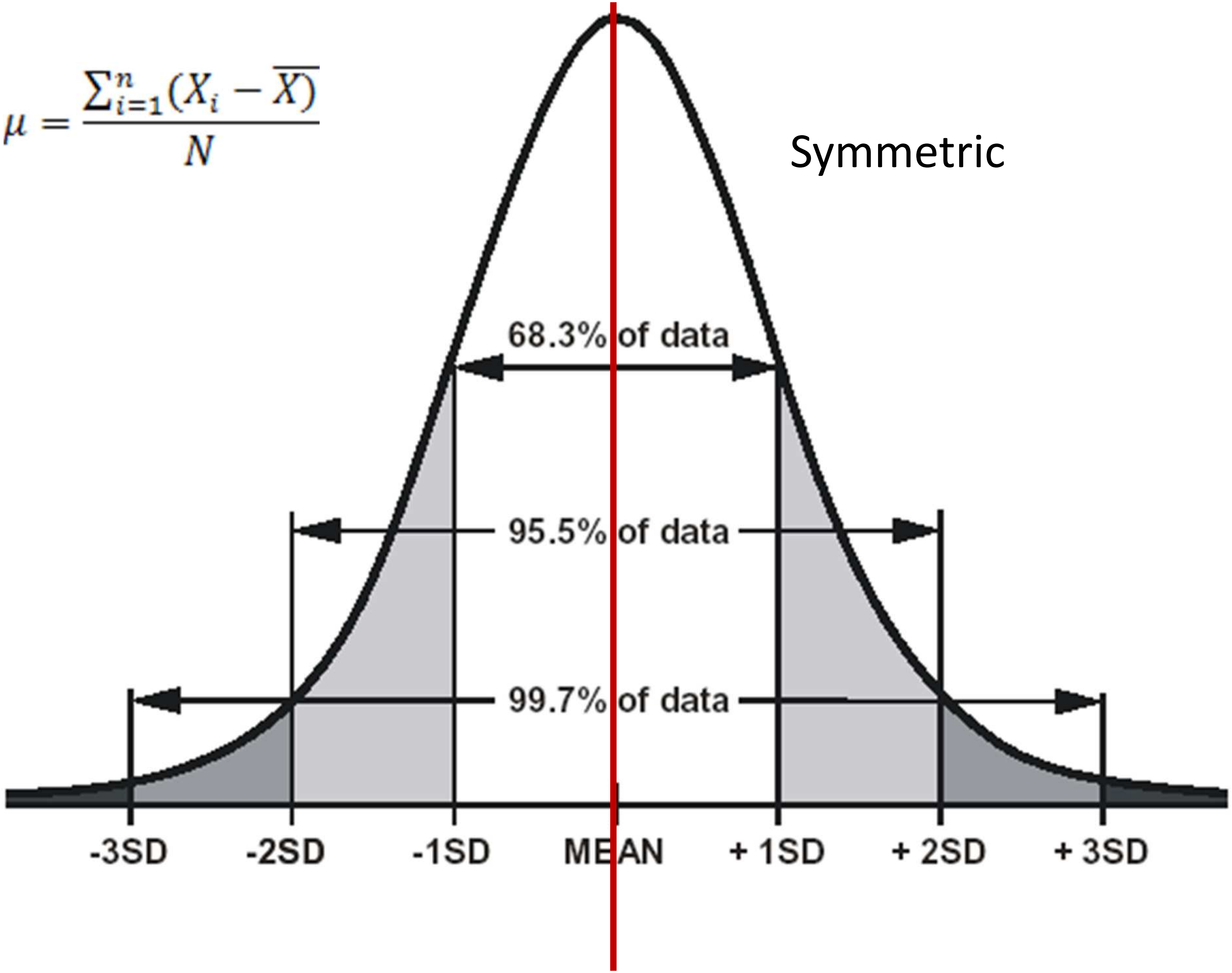
2.1. Normal distribution

- Symmetric about the mean: μ
- Mean = median = mode = μ, *SD* = σ
- Area under the curve = 1 implies 50% of the area to the left of the mean (or median)
- Notation: $X \sim N(\mu,\sigma^2)$
- *About 95% of the area is between* $\mu - 1.96\sigma$ *and* $\mu + 1.96\sigma$

Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean

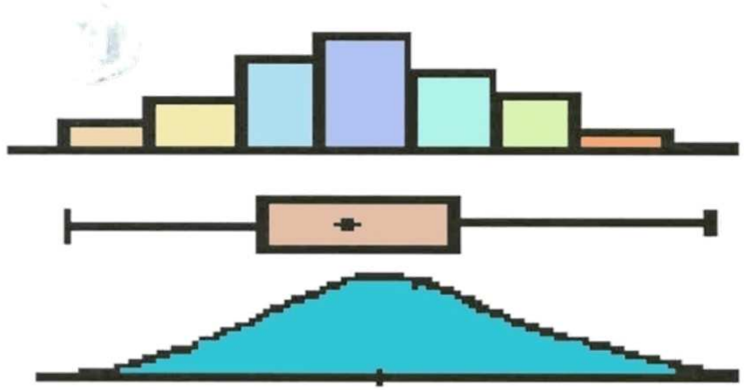
$$\mu = \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{N}$$

Symmetric



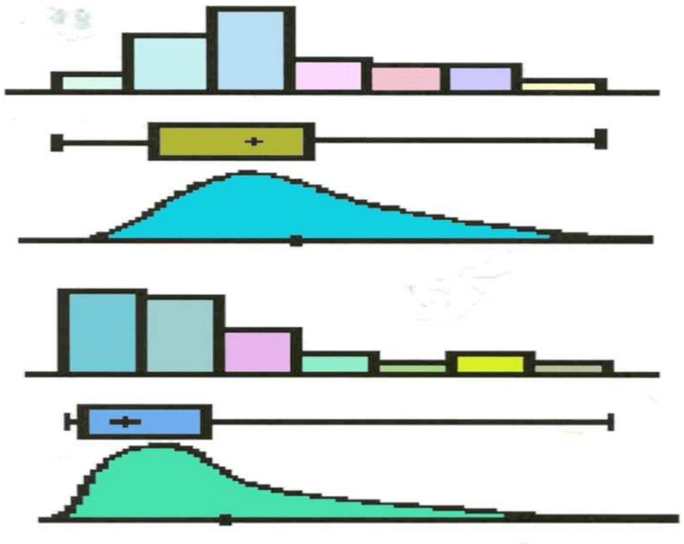
Shape of data distribution

- Normal Distribution

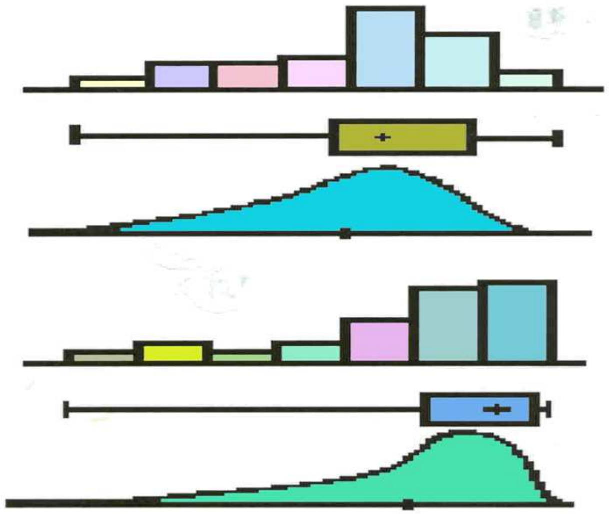


- Skewed Distribution

Positively skewed

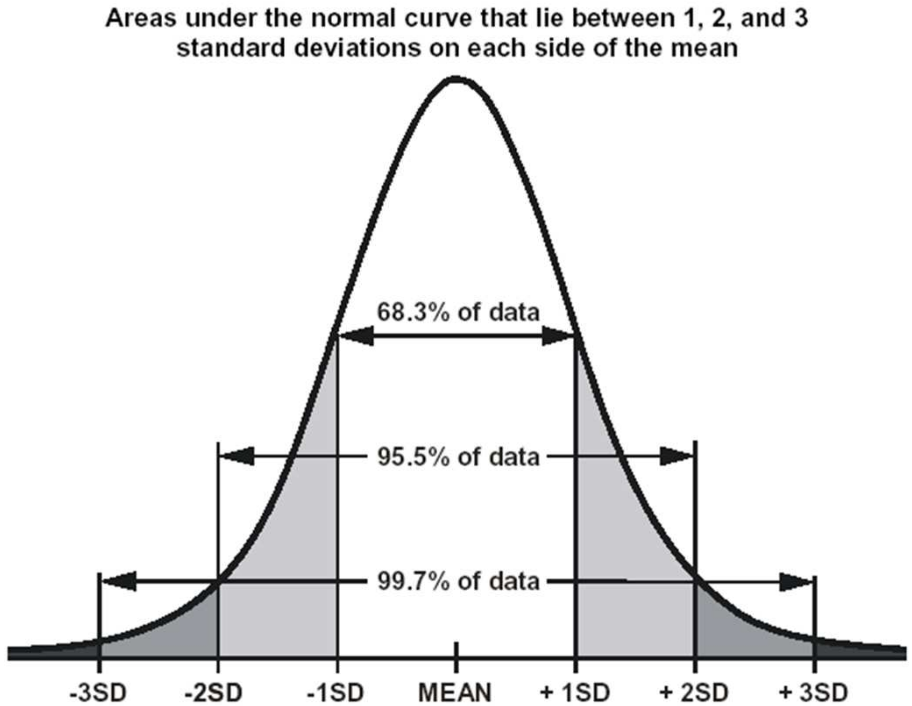


Negatively skewed



Standard normal distribution

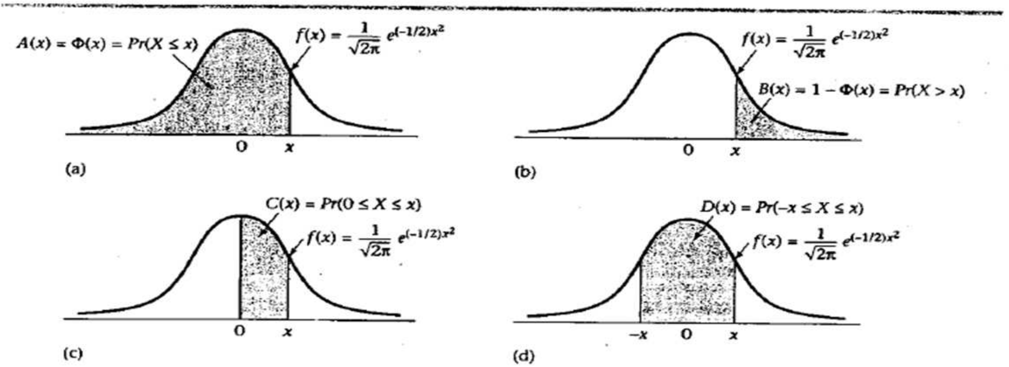
- Normal distribution
- Notation: $X \sim N(0,1)$
- Symmetric about mean = 0
- Standard deviation = 1



Standard normal distribution table

- See additional handout
- Column A : Area from - $\propto$(infinite) to *x*. This is the Pr ($X \leq x$)
- Column B : Area from *X to +*$\propto$. This is the Pr ($X \geq x$)
- Column C : Area from 0 to a point Pr ($0 \leq X \leq x$)
- Column D : Area symmetric about 0

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

Converting from a $N(\mu, \sigma^2)$ Distribution to a $N(0, 1)$ Distribution

- **The formula is:**

$$Z = \frac{X-\mu}{\sigma}$$

- ***Z* ~ *N*(0, 1)**

$$\text{Therefore,} \quad \Pr(a < X < b) = \Pr\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

Example

- μ = 200, σ^2 = 400 implies σ = 20, *a* = 220, *b* = 260

$$Z_1 = \frac{a-\mu}{\sigma} = \frac{220-200}{20} = 1.00$$

$$Z_2 = \frac{b-\mu}{\sigma} = \frac{260-200}{20} = 3.00$$

- We have converted the probability statement about *N*(200, 400) to one about *N*(0, 1). We can evaluate by using the standard normal tables.

$$\Pr(220 < X < 260) = \Pr\left(\frac{220-200}{20} < Z < \frac{260-200}{20}\right)$$

$$= \Pr(1.00 < Z < 3.00)$$

2.2. Binomial Distribution

- *n* independent trials, each with only two possible outcomes.
  - Example: Yes/No, Sick/not sick
- Probability of success at each trial is some constant, *p*
- *Probability of failure = 1 – p = q*

2.2. Binomial Distribution

- Probability of *k* successes in n trials is

$$\Pr(X = k) = \frac{n!}{(n-k)!k!} p^k q^{n-k}$$

  - $k = 0, 1, 2, \ldots, n$
  - $q = 1 - p$

  $(Pr = p)$

- $P(x \le n_a)$ = Sum of $Pr\ (x = n_i)$ with $i = 0, 1, \ldots n_a$
- $P(n_a \le x \le n_b)$ = Sum of $Pr\ (x = n_i)$ with $i = a, a+1, \ldots, b$

Normal approximation to the binomial

- When ***n* is large**, there are no tables and using the formula is difficult. The Normal Distribution can be used as an approximation to the **Binomial Distribution**

- The Normal Distribution is a good approximation **if *n* is large and *p* not too near 0 or 1**. The binomial then will be nearly symmetric when ***n*p*q*** $\geq$ **5**
- Binomial: mean = *np*, variance = *npq*. Let's create a normal distribution that is *N(np, npq)*
- If *n* is large nough, npq ≥ 5 --> X ~ B (np, npq), $\mu = n*p$; $\sigma^2 = n*p*q$

Normal approximation to the binomial

- We want Pr ($a \leq X \leq b$) where X is binomially distributed with parameters n and p. The first choice for an approximation is to examine the $N(np, npq)$ normal distribution by getting the area under this curve from a to b

- Continuity Correction: Better to get area under the normal curve
  - For X is between a and b: Pr ($a - ½ \leq X \leq b + ½$)
  - For X > a : Pr ($X \geq a - ½$)
  - For X < a : Pr ($X \leq b + ½$)

Example: operative complication

- Pr (at most 5 complications) = Pr ($X \leq 5$)
  - p = 0.20, n = 50
- Normal approximation:
  - mean = np = 10
  - variance = npq = 8 > 5
  - SD = Square root or SQRT (8) = 2.8284

$$\Pr(X \leq 5.5) = \Pr\left(Z \leq \frac{5.5 - 10}{2.8284}\right)$$

$$= \Pr(Z \leq -1.59)$$

$$= 1 - \Pr(Z \leq 1.59)$$

$$= 1 - 0.9441$$

$$= 0.0559$$

Thank you for your attention!