

## Index of 4-L

Page	Title
1	Practical information
2–3	Normal distributions
4	Exercise 1.109
5	Working with the normal: standardization
6	Details for standardization example
7	Normal plots
8	Introduction to normality tests
9	Example: Binomial distribution
10	Binomial setting
11	Exercises 5.32 and 5.33
12	Binomial distribution II
13	Statistical inference
14	Example: Parasite burdens in Lithuania
15	Example: Diagnostic testing
16	Distribution of a statistic
17	Summary notes

## PRACTICAL INFORMATION

WELCOME to all newly arrived students...!

### Scheduling news:

- do we want to schedule lab review sessions? and if yes, when? (Tuesdays 1-2pm or Wednesdays 1-2pm)
- final exam set for Monday, December 10, 9am-12pm,
- we also need to set the midterm exam around Oct 25th,
  - \* duration 1 hour,
  - \* exam is *optional*, but recommended to do it,
- first home assignment: will be posted next Thursday (on webpage and Moodle).

### Today's lecture:

- the normal distribution,<sup>1</sup>
- the binomial distribution<sup>2</sup> and *Poisson distribution*,<sup>3</sup>
- finally arrived at the real thing: statistical inference,<sup>4</sup>
  - \* parameters and statistics,
  - \* parameter estimation.

---

<sup>1</sup> PSLS 3e: Chapter 11; IPS7e: Section 1.3; S: Chapter 6.

<sup>2</sup> PSLS 3e: Chapter 12; IPS7e: Section 5.2; S: Chapter 5.

<sup>3</sup> Not a core course topic; PSLS Chapter 12.

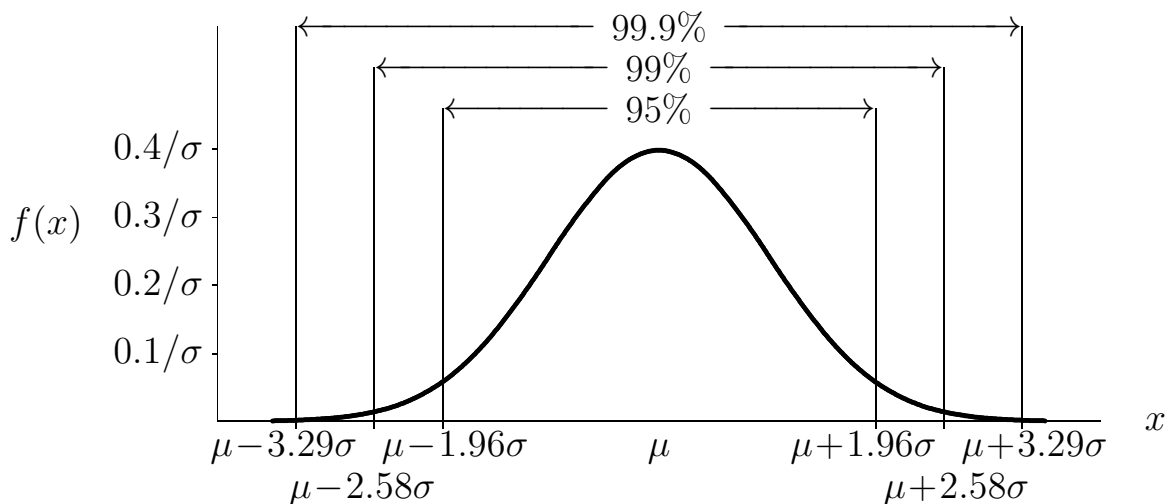
<sup>4</sup> PSLS 3e: Chapter 13 (part); IPS7e: Section 5.1 (part); S: Chapter 1 (part).

# NORMAL DISTRIBUTIONS

The normal distribution<sup>5</sup>, written in PSLS and IPS as  $N(\mu, \sigma)$ ,<sup>6</sup> is defined by its density curve:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where the mean  $\mu$  and standard deviation  $\sigma$  are the two parameters.



Comments and properties:

- “bell-shaped” (no matter  $\mu$  and  $\sigma$ ),
- symmetrical around  $\mu$  (= mean, median, mode),
- some more probabilities (the “68 – 95 – 99.7% rule”):

interval	$[\mu - \sigma, \mu + \sigma]$	$[\mu - 2\sigma, \mu + 2\sigma]$	$[\mu - 3\sigma, \mu + 3\sigma]$
probability	68%	95%	99.7%

<sup>5</sup> Also called the Gaussian distribution, after C. F. Gauss (1777-1855).

<sup>6</sup> It is more common to write  $N(\mu, \sigma^2)$  with the second parameter as the variance.

- unbounded (extends infinitely far out),
- probabilities not simple to calculate (see later),
- why so important?
  - \* mathematically very nice properties:
    - simple solutions to analysis-of-variance problems,
    - many “natural” distributions well fitted by normal due to certain mathematical laws,
  - \* works well, even without fitting the data perfectly,
  - \* historically important: before computers, the only distribution that enabled complicated analyses!

Standardized normal distribution  $N(0, 1)$  has  $\mu=0$  and  $\sigma=1$ ,

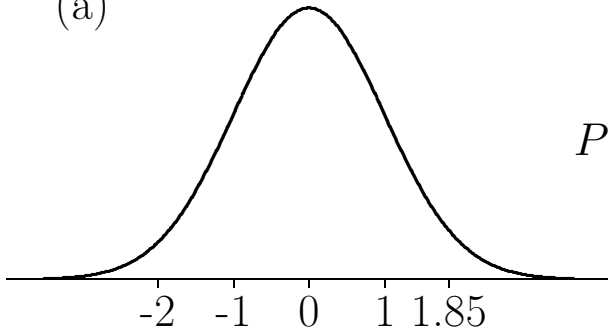
- all calculations in  $N(\mu, \sigma)$  can be translated to the standard normal (another special property),
- probabilities for  $N(0,1)$  stored in tables and computers:
  - \* typically  $P(Z < z)$  for different  $z$ -values, where  $Z$  is the usual name of a random variable from  $N(0,1)$ ,
  - \* tables: PSLS: Table B; S: Table 2; IPS: Table A,
  - \* Minitab: Calc-Prob.Distr.-Normal, choose entry Cumulative probability,<sup>7</sup>
  - \* also inverse probabilities (=percentiles) stored; in Minitab, choose instead entry Inverse cumulative probability.

---

<sup>7</sup> For a visual display, use the Graph-Probability Distribution Plot menu.

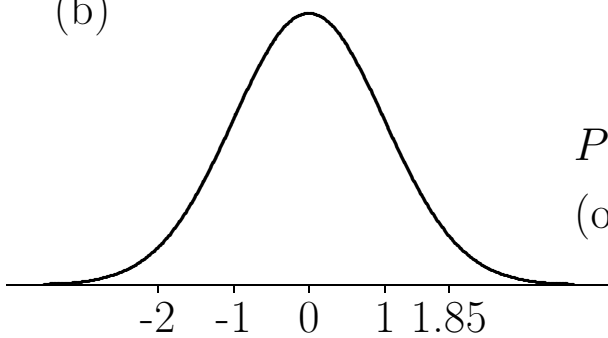
# EXERCISE 1.109

(a)



$$P(Z < 1.85) = 0.9678$$

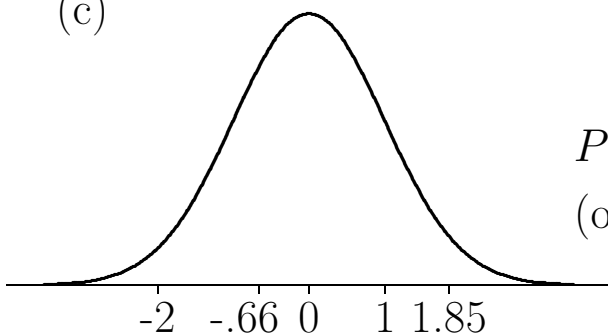
(b)



$$P(Z > 1.85) = 1 - 0.9678 = 0.0322$$

(or computed as  $P(Z < -1.85)$ )

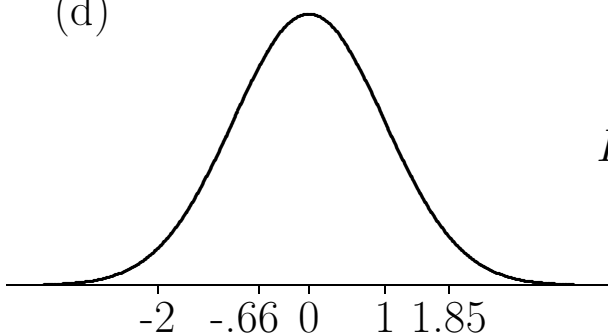
(c)



$$P(Z > -0.66) = 1 - 0.2546 = 0.7454$$

(or computed as  $P(Z < 0.66)$ )

(d)



$$\begin{aligned} P(-0.66 < Z < 1.85) &= P(Z < 1.85) - P(Z \leq -0.66) \\ &= 0.9678 - 0.2546 = 0.7132 \end{aligned}$$

## WORKING WITH THE NORMAL: STANDARDIZATION

Standardization is based on the following results

- (a) Any linear transformation  $x \mapsto y = a + bx$  transforms a random variable  $X \sim N(\mu_X, \sigma_X)$  into another normal variable  $Y = a + bX \sim N(\mu_Y, \sigma_Y)$ , where

$$\mu_Y (= EY) = a + b\mu_X, \quad \text{and} \quad \sigma_Y (= \text{sd}Y) = b\sigma_X.$$

- (b) In particular, for  $X \sim N(\mu_X, \sigma_X)$  we have

$$Z = (X - \mu_X) / \sigma_X \sim N(0,1),$$

and  $Z$  is the standardized form of  $X$ .

Normal distribution calculations (PSLS Example 11.6):

The lengths of human pregnancies from conception to birth (gestation period) follow roughly a normal distribution with  $\mu = 266$  and  $\sigma = 16$ . What proportion of babies are born after 240 days (8 months)?

We seek the probability, (where  $X \sim N(\mu, \sigma)$ )

$$P(X > 240) = P\left(\frac{X - 266}{16} > \frac{240 - 266}{16}\right) = P(Z > -1.625),$$

and table lookup gives  $P(Z > -1.63) = 1 - 0.0516 = 0.9484 = 94.8\%$ . More precisely, we could have used Minitab to get:  $P(Z > -1.625) = P(Z < 1.625) = 0.9479$ . (Note: the value  $z = (240 - \mu) / \sigma = -1.625$  is often called a z-score).

## DETAILS FOR STANDARDIZATION EXAMPLE

A more detailed derivation of the  $z$ -score goes as follows:

$$X > 240$$

or

$$X - 266 > 240 - 266$$

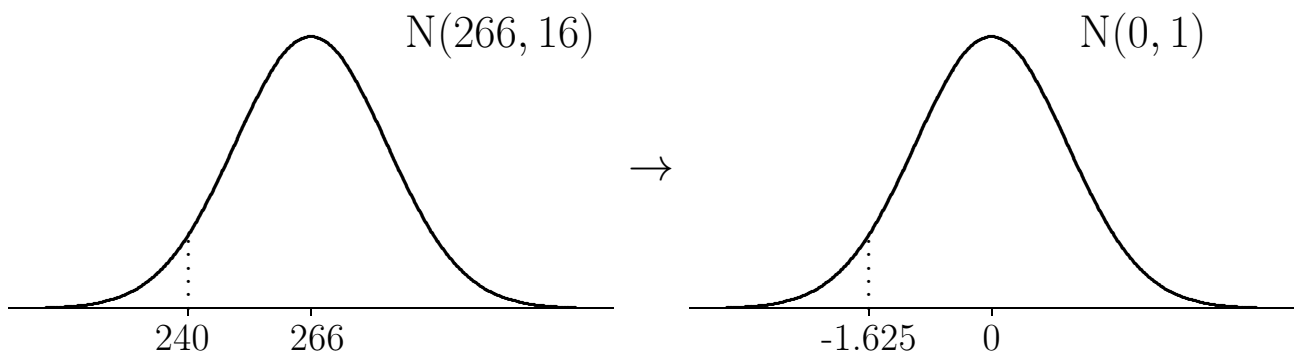
or

$$\frac{X - 266}{16} > \frac{240 - 266}{16}$$

or

$$Z > -1.625 \text{ (= } z\text{-score)}$$

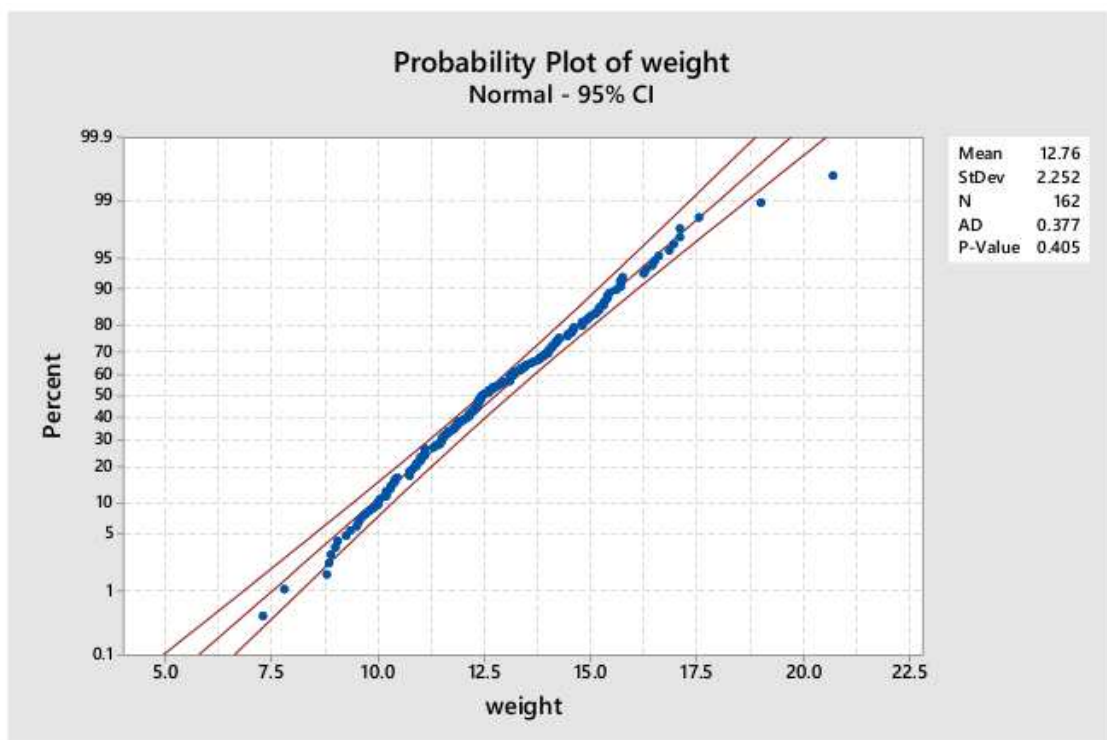
Graphically, we transfer the calculation from the left distribution to the right distribution in the figure below.



## NORMAL PLOTS

Normal quantile plots (normal probability plots<sup>8</sup>):

- graphical assessment of normality of dataset for a single variable,
- a straight line corresponds to a normal distribution (all values of  $\mu$  and  $\sigma$ ),
- informal assessment (tests used for formal assessments),
- if the plots looks bad, always make a histogram (or stem-plot) as well to more clearly see the problems.



Comment: reasonably straight line, with one point at the upper end clearly off, and others just beyond the confidence bands (indicating the “expected random variation”).

<sup>8</sup> The terms quantile plot and probability plot are used interchangeably, but usually a probability plot has percentages or  $z$ -scores on the  $y$ -axis, whereas a quantile plot usually has  $z$ -scores on the  $x$ -axis.

## INTRODUCTION TO NORMALITY TESTS

Main idea (more details about statistical tests later):

the  $P$ -value indicates data's agreement with a normal distribution:  
(“null hypothesis  $H_0$ ”: distribution *is* normal)

- low  $P$ -value: data do not appear normal,
- high  $P$ -value: data may very well be normal,
- cut-off for low/high:  $P=0.05$  most commonly used.

Implementation:

- calculation utilizes statistical software, e.g. Minitab:  
Basic Stats-Normality Test (3 tests),<sup>9</sup>
- many tests exist, and they usually agree reasonably well.

Why of interest to “test for normality”...?

- if a normal distribution approximates well, we may use it for probability calculations,
- many methods for statistical inference involve assumed normal distributions  $\Rightarrow$  use test as a validation step.

Caution: — it's one of the most misused statistical procedures... ,

- observations must be from a single distribution, not several ones pooled together (e.g., females and males),
- many methods for statistical inference do not assume the raw data values to be normally distributed (instead the “residuals”),
- mild violations of normality may not be a real problem.

---

<sup>9</sup> The Stata commands `sktest`, `swilk` or `sfrancia` give 3 different tests; R has the command `shapiro.test`.

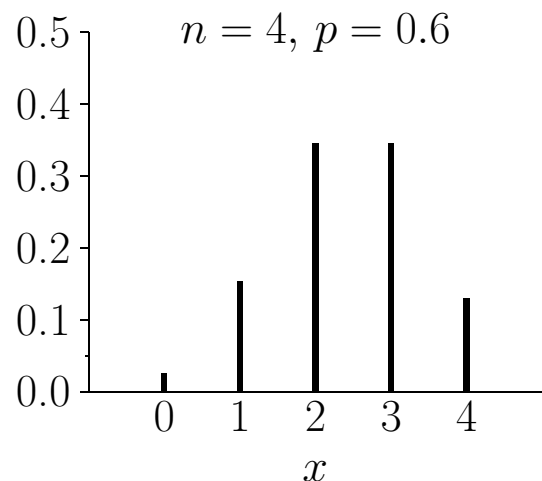
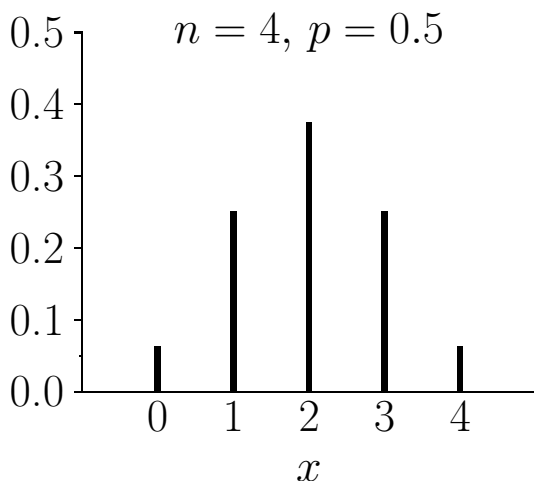
EXAMPLE: BINOMIAL DISTRIBUTION

Assume we toss a coin 4 times, and that

- the coin has the same probability  $p$  to show heads (H) in every run,
- the outcomes of the 4 runs are independent.

Then the total number of heads in the 4 runs follows a binomial distribution  $B(n, p)$  with denominator  $n = 4$  and probability  $p$ . The probabilities are,

- $p(4) = P(\text{“sequence HHHH”}) = p \times p \times p \times p = p^4,$
- $p(0) = P(\text{“sequence TTTT”}) = (1 - p)^4,$
- $p(3) = P(\text{“sequence HHHT, HHTH, HTHH or THHH”}) = 4p^3(1 - p),$
- $p(1) = 4p(1 - p)^3,$
- $p(2) = 6p^2(1 - p)^2.$



## BINOMIAL SETTING

The binomial setting involves the following assumptions,

- a fixed number  $n$  of observations (often trials),
- the  $n$  observations are all independent,
- each observation takes one of two possible values (categories) — “success” (1) or “failure” (0),
- the probability of “success” is the same,  $p$ , for all observations.

⇒ a binomial distribution  $(n, p)$  for number of successes.

Typical examples: outcomes such as germination, survival, presence/absence of certain phenomena (especially disease), answer yes/no to questions, etc.

Sampling of  $n$  elements from a finite population ( $N$  elements) with two categories of elements (of which proportion  $p$  is of first category),

- with replacement ⇒ binomial setting  $(n, p)$ ,
- without replacement ⇒ approximate binomial setting when  $n \ll N$ ; rule of thumb for use of approximation is that  $N \geq 20n$ .

Sampling without replacement from a finite population is modelled (exactly) by a hypergeometric distribution (not covered in our textbooks).

EXERCISES 5.32 AND 5.33
-------------------------

Using a binomial distribution.

Exercise 5.32:

- (a) A binomial distribution  $B(50, p)$ , where  $p =$  prob. of girl, seems reasonable, under one assumption about the data collected (which?).
- (b) Clearly not a binomial distribution. (Why not?)
- (c) It's probably not reasonable to assume a binomial distribution  $B(50, p)$ . (Why not?)

Exercise 5.33:

- (a) A binomial distribution  $B(50, p)$ , where  $p =$  prob. of passing the exam. One should consider which population the students are intended to represent, and also whether they were taught together or individually.
- (b) Not a binomial distribution  $B(10, p)$ . (Why not?)
- (c) Not a binomial distribution  $B(10, p)$ . (Why not?)

## BINOMIAL DISTRIBUTION II

Binomial distribution  $(n, p)$  (or  $B(n, p)$  or  $\text{Bin}(n, p)$ ), corresponding to a binomial setting:

- $n$  = number of trials/observations (denominator),
- $p$  = probability parameter,  $0 \leq p \leq 1$ ,

has the following properties,

- the probability function  $p(x)$  is given by

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \dots, n,$$

where the binomial coefficient  $\binom{n}{x} = \frac{n!}{x!(n-x)!}$  is the number of ways to select a subset of  $x$  elements from a set of  $n$  elements (“ $n$  choose  $x$ ”), and  $n!$  is the factorial<sup>10</sup>,

- mean  $\mu = np$ ,
- standard deviation  $\sigma = \sqrt{np(1-p)}$ .

Probabilities in  $B(n, p)$  can be obtained by

- calculation by hand (calculator),
- tables: PSLS: no table; S: Table 1; IPS: Table C — for selected values of  $(n, p)$ ,
- Minitab<sup>11</sup> menu: Calc-Prob.Distrib.-Binomial,
- approximations (next lecture).

---

<sup>10</sup> Factorial function:  $n! = 1 \cdot 2 \cdot \dots \cdot (n-1) \cdot n$ , for integers  $n \geq 1$ , and  $0! = 1! = 1$ .

<sup>11</sup> The Stata function `binomial(n, k, p)` gives  $P(X \leq k)$  for  $X \sim B(n, p)$ ; in R, the function is: `pbinom(k, n, p)`.

## STATISTICAL INFERENCE

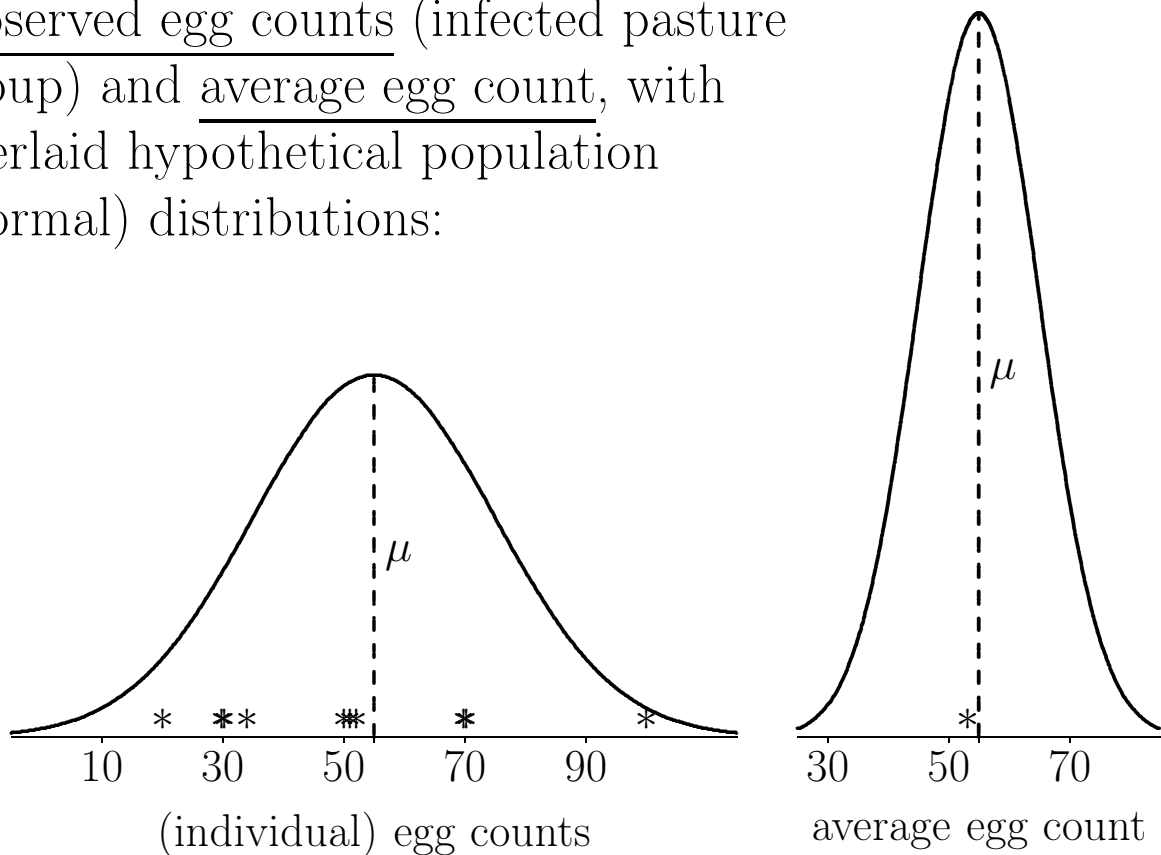
- to draw conclusions from data about some wider population (for which the data are representative),
- relies always on assumptions of statistical models,
- procedures (methods, algorithms, programs) developed to supplement, *not substitute* our common sense.

### Outline of statistical analysis (revisited):

- Data description.
- Statistical models: we formulate statistical models involving probability distributions and parameters (e.g.,  $\mu$ ):
  - \* fixed, unknown numbers describing a population,
  - \* the primary focus of statistical inference.
- Estimation: our information about the parameters from the observed data is summarized in statistics or estimates:
  - \* quantities calculated from the data to give possible parameter values, e.g. the sample mean  $\bar{X}$ ,
  - \* standard notation: parameters with a “hat”, e.g.  $\hat{\mu} = \bar{X}$ ,
  - \* aim of statistical methodology: obtain estimates as close to true values as possible (though *rarely* (never) equal to true values),
  - \* any estimate should be accompanied by a measure of its uncertainty: e.g. its standard deviation (often called *standard error*, SE or SEM) or a confidence interval.

## EXAMPLE: PARASITE BURDENS IN LITHUANIA

Observed egg counts (infected pasture group) and average egg count, with overlaid hypothetical population (normal) distributions:



Interpretations:

- population: calves on infected pasture, under similar conditions as in spring-summer in Lithuania,
- model: sample of size 10 from  $N(\mu, \sigma)$ ,
- parameters:  $\mu$  and  $\sigma$  = population mean and standard deviation of egg counts after 10 weeks on pasture,
- estimates:

$$\hat{\mu} = \bar{X} = 51.2 \quad (\text{sample average}),$$

$$\hat{\sigma} = s = 24.0 \quad (\text{sample standard deviation}).$$

## EXAMPLE: DIAGNOSTIC TESTING

Assume a new laboratory diagnostic test under development for detection of a disease; e.g. ISA (infectious salmon anemia).

Two essential characteristics of the test:

- *sensitivity* = proportion of true positive animals detected by the test, (e.g.,  $Se = 100\% \Rightarrow$  no false negatives),
- *specificity* = proportion of true negative animals declared negative by the test, (e.g.,  $Sp = 100\% \Rightarrow$  no false positives).

Interpretations (sensitivity):

- population: “truly positive animals” — must be defined by some other criterion/test (maybe a “gold standard”),
- model: each infected animal has the same probability  $p$  of testing positive by the test  $\Rightarrow$  Binomial distribution  $(n, p)$ ;
  - \* do population characteristics such as age, sex, severity of disease etc., affect the probability of testing positive?
- parameter:  $p$  (sensitivity),
- estimate:  $\hat{p}$  = proportion of positives in sample.

In practice, sensitivity and specificity calculations are limited to well-defined and often fairly narrow populations.

## DISTRIBUTION OF A STATISTIC

Let our data be  $X_1, \dots, X_n$  (e.g., the parasite egg counts).

More formally about estimates:

- estimates (in fact, all statistics) are functions of the data:

$$X_1, \dots, X_n \mapsto f(X_1, \dots, X_n),$$

for example, the average  $\bar{X} = (X_1 + \dots + X_n)/n$ ,

- estimates are random variables and have distributions;  
— *intuitively*, if the  $X$ 's are random, so is the average,
- the distribution of estimates depends on the unknown parameter(s) of the model;  
— *intuitively*, if the distribution of the  $X$ 's does, the same must be true for statistics computed from them.

But ... where does the variation come from?

- sometimes sampling variability, when randomly selecting a sample from a finite population (different members of the population get selected),
- sometimes physical variation, like measurement error, laboratory variation etc.,
- sometimes biological variation, when “homogeneous” experimental units (e.g. animals) react differently upon experimental conditions.

## SUMMARY NOTES

The normal distribution  $N(\mu, \sigma)$ :

- parameters:  $\mu$  (mean) and  $\sigma$  (standard deviation),
- symmetry, “bell-shape”, 68 – 99 – 99.7 rule,
- standard normal ( $N(0, 1)$ ,  $z$ -distribution), and methods for getting probabilities/percentiles,
- standardization,  $z$ -score,
- normal quantile/probability plot, normality test.

The binomial distribution  $B(n, p)$ :

- parameters  $n$  (no. trials) and  $p$  (probability),
- binomial setting, approx. of simple random sampling,
- probability function, mean, standard deviation.

The Poisson distribution with mean  $\lambda$  for counts of events<sup>12</sup>

- probability function:  $p(x) = e^{-\lambda} \lambda^x / x!$ , for  $x = 0, 1, 2, \dots$

Key words and concepts in statistical inference:

- parameter, estimate, population,
- distribution of estimate/statistic,
- statistical model (a formal set of assumptions needed to make the inference valid).

---

<sup>12</sup> Counts with no natural upper bound and sample space  $S = \{0, 1, 2, \dots\}$ ; typical examples: traffic accidents, cases of a non-infectious disease, bacteria colonies on a plate, plants in an area.