

Index of 4-L: Normal and Binomial Distributions

Page	Title
1	Practical information
2–3	Normal distributions
4	Exercise 1.109
5	Working with the normal distribution: 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

Scheduling and organization:

- **first home assignment**: will be posted next Thursday (on webpage and Moodle),
- we need to set the **final exam**: April 20th or move up to April 17th?,
- we also need to set the **midterm exam** around March 5th,
 - * duration 1 hour,
 - * midterm exam is *optional*, but recommended to take it.

Today's lecture:

- the **normal distribution**,¹
- the **binomial distribution**² and Poisson distribution³,
- finally arrived at the real thing: **statistical inference**,⁴
 - * parameters (introduced on 3L–16) and statistics,
 - * the first component of inference: **parameter estimation**.

¹ PSLS 4e: Chapter 11; IPS7e: Section 1.3; S: Chapter 6.

² PSLS 4e: Chapter 12; IPS7e: Section 5.2; S: Chapter 5.

³ Not a core course topic; PSLS 4e: Chapter 12, IPS 9e: Section 5.3.

⁴ PSLS 4e: Chapter 13 (part); IPS7e: Section 5.1 (part); S: Chapter 1 (part).

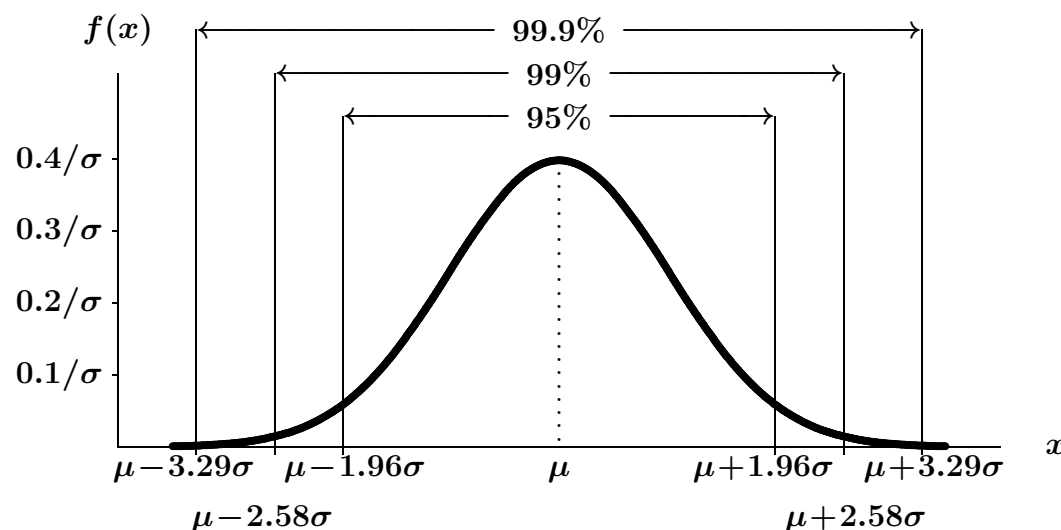
NORMAL DISTRIBUTIONS

The **normal distribution**⁵, written in PSLS and IPS as $N(\mu, \sigma)$,⁶ is defined by its **density function/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 μ and standard deviation σ are the two **parameters**:

- “bell-shaped”
(no matter μ and σ),
- **symmetrical** around μ :
the mean, median, and mode,
- **unbounded**
(extends infinitely far out),
- 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%

- probabilities not simple to calculate (more on the following pages),

⁵ Also called the **Gaussian** distribution, after C. F. Gauss (1777-1855).

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

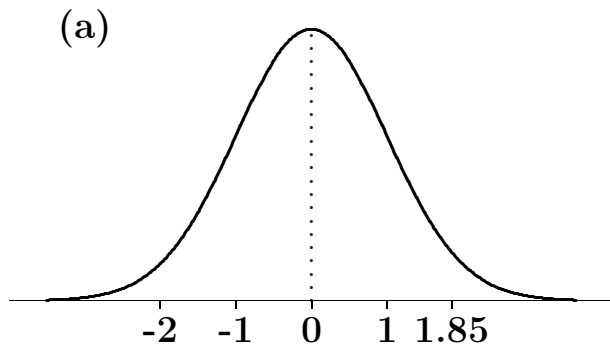
- why is the normal distribution **so important?**
 - * mathematically very nice properties:
 - gives simple solutions to analysis-of-variance problems,
 - many “natural” distributions are well fitted by normal due to certain mathematical laws (next lecture),
 - * **works well**, even without fitting the data perfectly,
 - * **historically important**: before computers, it was the only distribution that enabled complex analysis!

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

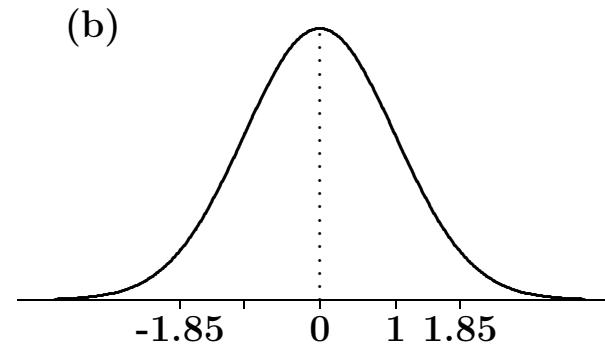
- all calculations in $N(\mu, \sigma)$ can be translated to the standard normal $N(0, 1)$ (another special property),
- **probabilities for $N(0, 1)$** are stored in tables and computers:
 - * typically $P(Z < z)$ for different z -values⁷, where Z is the usual name of a random variable from $N(0, 1)$,
 - * **tables** — PSLS: Table B; S: Table 2; IPS: Table A,
 - * Minitab **“table”**: Calc-Prob.Distribution-Normal, choose: Cumulative probability; also **inverse probabilities** (=percentiles), at Inverse cumulative probability,
 - * Minitab **visual displays**: in the menu Graph-Probability Distribution Plot.

⁷ $P(Z < z)$ is called a **cumulative probability**, or (Henrik) “probability to the left”.

EXERCISE 1.109

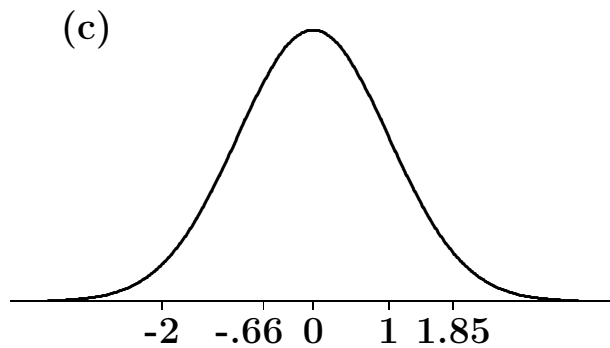


$$P(Z < 1.85) = 0.9678$$



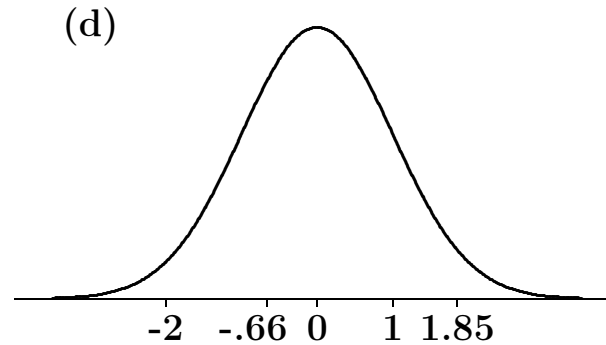
$$P(Z > 1.85) = 1 - 0.9678 = 0.0322$$

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



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

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



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

WORKING WITH THE NORMAL DISTRIBUTION: 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

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

- (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 3e Example 11.6/11.7):

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

$$X - 266 > 240 - 266$$

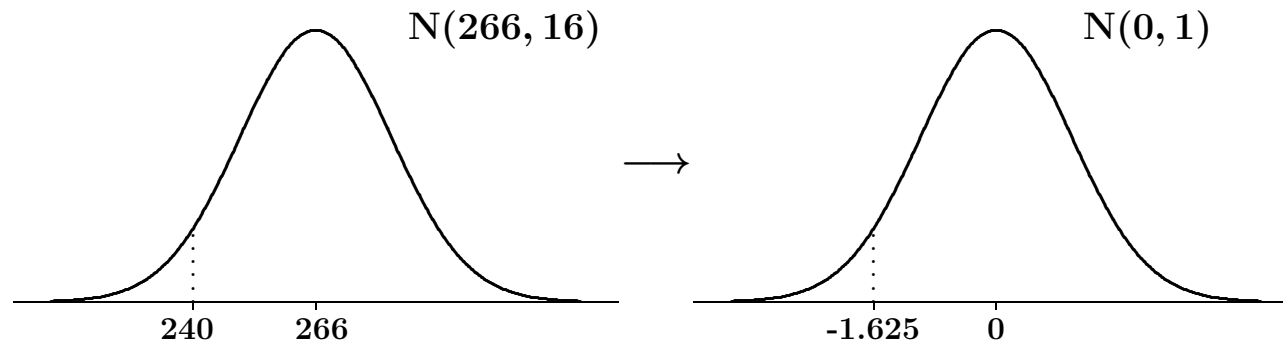
or

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

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.

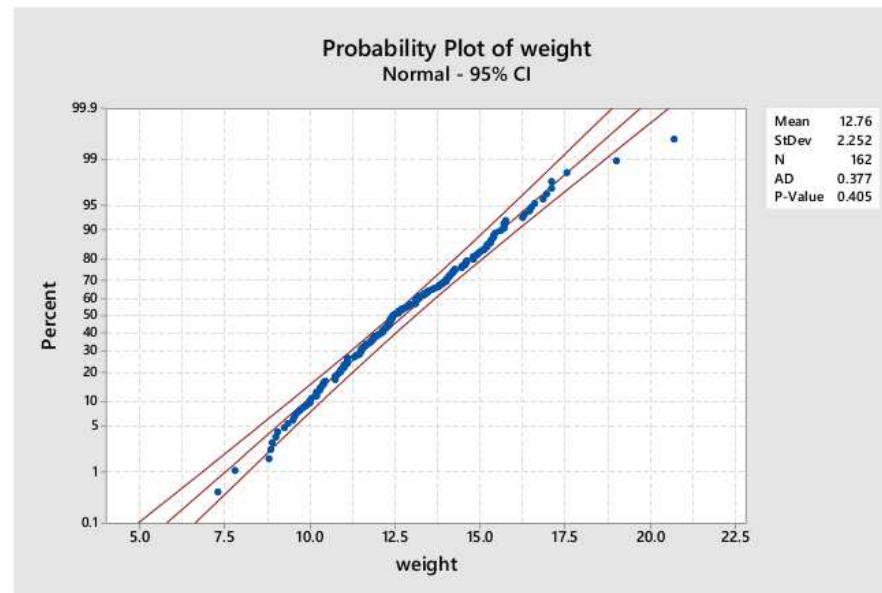


NORMAL PLOTS

Normal quantile plots (normal probability plots⁸):

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

Crab data:
(from 1–L)



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”).

⁸ 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 the data's agreement with a normal distribution, more precisely⁹

- **low P -value**: data do not appear normal; **high P -value**: data may very well be normal,
- definitive cut-off for low/high, such as: low $\sim P \leq 0.05$, usually not needed (the interpretation depends on the sample size, and on what the assessment of normality is used for).

Implementation:

- calculation utilizes statistical software, e.g. Minitab: Basic Stats-Normality Test (3 tests),¹⁰
- **many different 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
⇒ use test as a **validation** step for the inference.

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

- the 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 for the validity of results.

⁹ In the statistical testing framework, the “**null hypothesis H_0** ” is: distribution *is* normal.

¹⁰ The Stata commands: `swilk`, `sktest` and `sfrancia` give 3 different tests; R has the function: `shapiro.test`.

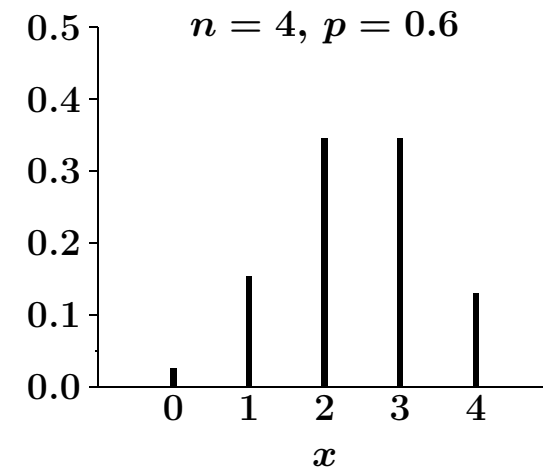
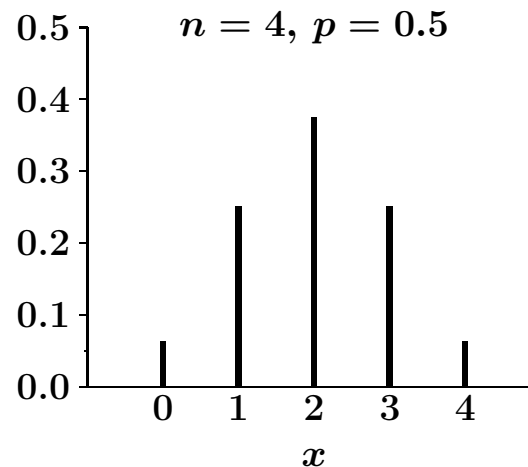
EXAMPLE: BINOMIAL DISTRIBUTION

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/trials 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,

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



BINOMIAL SETTING

A **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 $B(n, p)$** for the number of successes (out of the total n).

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

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

- (*rare*): **with replacement** ⇒ binomial setting (n, p),
- (*in practice*): **without replacement** ⇒ approximate binomial setting when $n \ll N$;
rule of thumb for use of approximation is that $N \geq 20n$ (or, $n/N \leq 5\%$).

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 $p =$ probability of girl, seems reasonable, under one assumption about the data collected. (which? — no twins)
- (b) Clearly not a binomial distribution. (why not? — no fixed n)
- (c) It's probably not reasonable to assume a binomial distribution $B(50, p)$. (why not? — dependence within a couple, not same probability for women and men)

Exercise 5.33:

- (a) A binomial distribution $B(50, p)$, where $p =$ probability 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? — because of the instruction between questions, p is not constant)
- (c) Not a binomial distribution $B(10, p)$. (why not? — p not constant, due to the different temperatures)

BINOMIAL DISTRIBUTION II

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

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

has **probability function** $p(x)$,

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

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**¹¹,

- **mean** $\mu = np$, and **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) ,
- Minitab¹² “**table**” menu: Calc-Probability Distribution-Binomial,
- **approximation** formulas (next lecture).

¹¹ Factorial function: $n! = 1 \cdot 2 \cdot \dots \cdot (n-1) \cdot n$, for integers $n \geq 1$, and $0! = 1! = 1$.

¹² 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

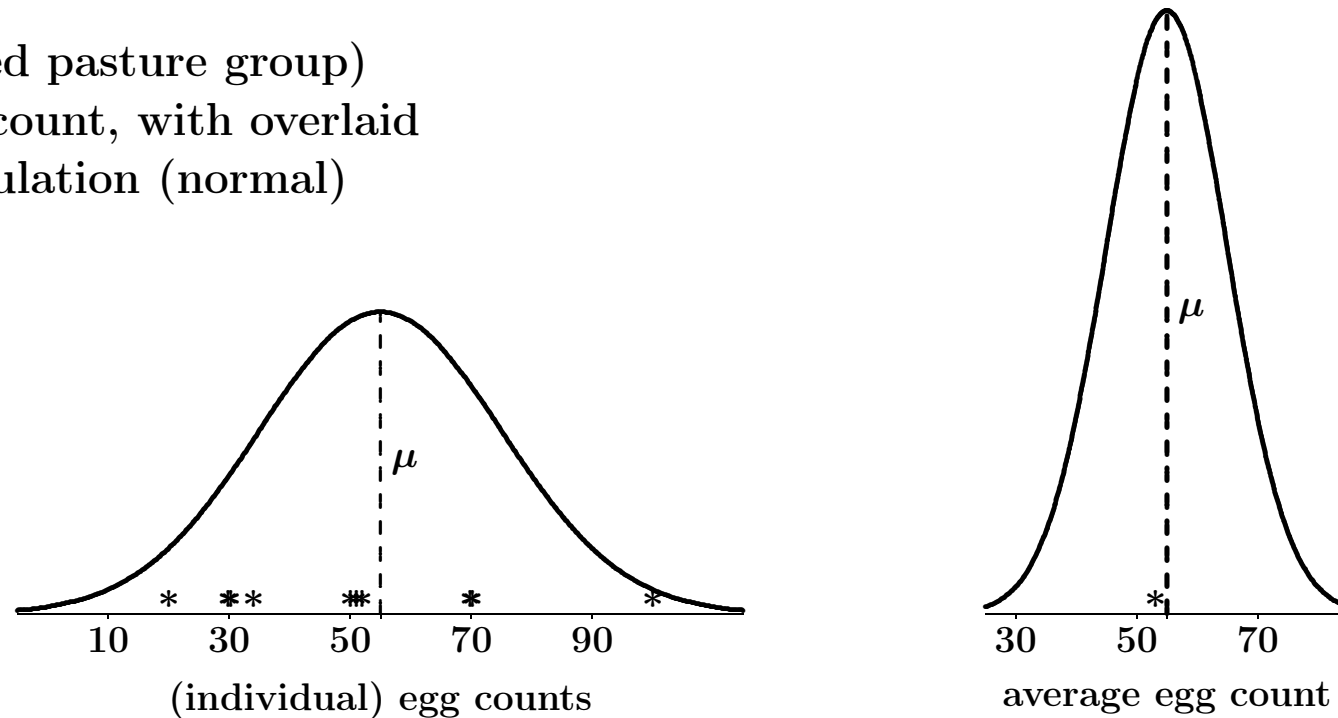
- to draw conclusions from data about some wider population (for which the data are representative),
- relies always on assumptions of statistical models,
- methods developed to supplement, **not substitute**, our common sense.

Outline of statistical analysis:

- **Data description** (Descriptive statistics).
- **Statistical models**: we formulate statistical models involving theoretical probability distributions and their **parameters** (e.g., μ):
 - * 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**:
 - * 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 a **confidence interval**.

EXAMPLE: PARASITE BURDENS IN LITHUANIA

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



- **population:** calves on infected pasture, under similar conditions as in spring-summer in Lithuania,
- **model:** sample of size 10 from $N(\mu, \sigma)$,
- **parameters:** μ and σ = population mean and standard deviation of egg counts after 10 weeks on pasture,
- **estimates:**
 - $\hat{\mu} = \bar{X} = 51.2$ (sample average),
 - $\hat{\sigma} = s = 24.0$ (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,¹³
- **specificity** = proportion of **true negative** animals declared negative by the test.¹³

Interpretations (sensitivity):

- **population**: “true 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** $B(n, p)$;
 - * 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.

¹³ We already (Lecture 3) encountered sensitivity (Se) and specificity (Sp); recall that $Se = 100\% \Rightarrow$ no false negatives, and $Sp = 100\% \Rightarrow$ no false positives.

DISTRIBUTION OF A STATISTIC

Let our data be (random variables) 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 population (because 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: μ (mean) and σ (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 (number of trials) and p (probability),
- binomial setting, approximation for simple random sampling: 5% rule,
- probability function, mean, standard deviation.

The **Poisson distribution** with mean λ for counts of events¹⁴

- 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).

¹⁴ 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.