5 Statistics

(5.1) Definition. A *random sample* of length $n \in \mathbb{N}$ is a set of n independent, identically distributed (iid) random variables $X_1, \ldots, X_n$.

(5.2) Remark. The data sets in section 1 can be considered as concrete representations of measurements of n random experiments, which are given by samples $\omega_i \in \Omega$, where Ω denotes a (common) sample space underlying the definition of X_i.
Alternatively, the data sets can be considered as the composed output of an experiment with underlying sample space Ω^n.

(5.3) Definition. A *statistic* is a function $\mathcal{S} : \bigcup_{n\in\mathbb{N}} \mathbb{R}^n \to \mathbb{R}$, which can be applied to any random sample (or any finite data set).

(5.4) Remark. The word *statistics* is not only used to denote the plural of a statistic in the above sense, but rather refers to the scientific discipline of the collection, presentation, analysis and interpretation of data.

With respect to random samples the science of statistics provides methods to draw conclusions about the underlying distribution from a sampled data set. In particular, if the underlying distribution is specified up to a set of unknown parameters, the aim of *parametric statistical inference* is to obtain estimations of the unknown parameters from sampled data sets.

Non-parametric statistical inference describes methods, that may be applied, if even the underlying distribution is completely unknown.

(5.5) Definition. A statistic that is used to estimate a parameter of a distribution underlying a random sample is called an *estimator*.

If the expectation of an estimator (as a function of the random variables $X_1, \ldots, X_n$ of the random sample) is equal to the estimated parameter, the estimator is said to be *unbiased*.

(5.6) Definition. The *sample mean* of a random sample $X_1, \ldots, X_n$ is defined to be:

$$\overline{X} \;:=\; \frac{X_1 + \cdots + X_n}{n}$$

(5.7) Lemma. Let $X_1, \ldots, X_n$ be a random sample with $\mu = E(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$. Then:

(i) $E(\overline{X}) = \mu$, i. e. $\overline{X}$ is an unbiased estimator for $E(X_i)$

(ii) $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$

(5.8) Lemma. For a random sample with $\mu = E(X_1)$, $\sigma^2 = \mathrm{Var}(X_1)$ and sufficiently large n, the central limit theorem (4.54) implies:

$$\Pr\left(|\overline{X} - \mu| < d\right) \approx p \qquad \Longleftrightarrow \qquad d \approx \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)$$

Proof.

$$\begin{aligned} p \;=\; \Pr\left(|\overline{X}-\mu| \;<\; d\right) \;&=\; \Pr\left(\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \;<\; \frac{d}{\sigma/\sqrt{n}}\right) \\ &=\; \Pr\left(-\frac{d}{\sigma/\sqrt{n}} \;<\; \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \;<\; \frac{d}{\sigma/\sqrt{n}}\right) \\ &\approx\; \Phi\left(d\cdot\frac{\sqrt{n}}{\sigma}\right) \;-\; \Phi\left(-d\cdot\frac{\sqrt{n}}{\sigma}\right) \;=\; 2\cdot\Phi\left(d\cdot\frac{\sqrt{n}}{\sigma}\right) \;-\; 1 \end{aligned}$$

$$\iff \qquad \frac{1+p}{2} \;\approx\; \Phi\left(d\cdot\frac{\sqrt{n}}{\sigma}\right)$$

$$\iff \qquad d \;\approx\; \frac{\sigma}{\sqrt{n}}\cdot\Phi^{-1}\left(\frac{1+p}{2}\right)$$

□

(5.9) Definition. The *sample variance* and *sample deviation* of a random sample $X_1,\ldots,X_n$ are defined to be:

$$S^2 \;:=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \qquad \text{and} \qquad S \;:=\; \sqrt{S^2}$$

respectively, where $\overline{X}$ is the sample mean of the random sample.

(5.10) Lemma. Let $X_1,\ldots,X_n$ be a random sample with $\mu = E(X_1)$ and $\sigma^2 = \mathrm{Var}(X_1)$. Then S^2 is an unbiased estimator for σ^2:

$$E(S^2) \;=\; \sigma^2$$

5.1 Random samples from a normal distribution

(5.11) Theorem. Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim \mathcal{N}(\mu, \sigma)$. Then $\overline{X}$ and S^2 are independent with:

$$\overline{X} \;\sim\; \mathcal{N}(\mu, \sigma/\sqrt{n}) \qquad \text{and} \qquad \frac{n-1}{\sigma^2} S^2 \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \;\sim\; \chi^2_{n-1}$$

(5.12) Corollary. Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim \mathcal{N}(\mu, \sigma)$. Then:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \;\sim\; t_{n-1}$$

5.1.1 Confidence intervals for μ

(5.13) Notation. Let $\overline{x}$ denote the value of the sample mean of a random sample. Given the construction of intervals $I = I(\overline{x})$ for all such values $\overline{x}$, such that

$$\Pr(\mu \in I) \;=\; 1 - \alpha$$

for a fixed $\alpha \in (0, 1)$, the intervals are called *confidence intervals* of *confidence level* $1 - \alpha$ for μ.

If I is a symmetric *two-sided confidence interval* with respect to $\overline{x}$, i.e.

$$I \;=\; [\overline{x} - \delta, \overline{x} + \delta]$$

for some $\delta \in \mathbb{R}^+$, this may be stated as:

$$\mu \;=\; \overline{x} \pm \delta \qquad \text{with a confidence of } \; 100(1-\alpha)\,\%$$

(5.14) Determination of confidence intervals for μ, if σ^2 is known.

Using the sample mean $\overline{X}$ to estimate μ, it follows that:

(i) Two-sided confidence intervals

$$1 - \alpha \;=\; \Pr\Big(\mu \;\in\; [\overline{X} - \delta,\, \overline{X} + \delta]\Big)$$

$$\iff \qquad 1 - \alpha \;=\; \Pr\Big(|\overline{X} - \mu| \;\leq\; \delta\Big) \;=\; 2\,\Phi\left(\frac{\delta}{\sigma/\sqrt{n}}\right) - 1 \qquad \text{(see proof of (5.8))}$$

$$\iff \qquad 1 - \frac{\alpha}{2} \;=\; \Phi\left(\frac{\delta}{\sigma/\sqrt{n}}\right)$$

$$\iff \qquad \delta \;=\; \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Therefore, with a confidence level of $1 - \alpha$:

$$\mu \;=\; \overline{X} \;\pm\; \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

(ii) One-sided upper confidence intervals

$$1-\alpha \;=\; \Pr\Big(\mu \in [\overline{X}-\delta,\infty)\Big) \;=\; \Pr\Big(\mu \geq \overline{X}-\delta\Big)$$

$$\iff \qquad 1-\alpha \;=\; \Pr\Big(\overline{X}-\mu \leq \delta\Big) \;=\; \Pr\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{\delta}{\sigma/\sqrt{n}}\right) \;=\; \Phi\left(\frac{\delta}{\sigma/\sqrt{n}}\right)$$

$$\iff \qquad \delta \;=\; \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}(1-\alpha)$$

Therefore, with a confidence level of $1-\alpha$:

$$\mu \;\geq\; \overline{X} \;-\; \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}(1-\alpha)$$

(iii) One-sided lower confidence intervals

$$1-\alpha \;=\; \Pr\Big(\mu \in (-\infty,\overline{X}+\delta]\Big) \;=\; \Pr\Big(\mu \leq \overline{X}+\delta\Big)$$

$$\iff \qquad 1-\alpha \;=\; \Pr\Big(-\delta \leq \overline{X}-\mu\Big) \;=\; \Pr\left(-\frac{\delta}{\sigma/\sqrt{n}} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)$$

$$= \;1 \;-\; \Phi\left(-\frac{\delta}{\sigma/\sqrt{n}}\right) \;=\; \Phi\left(\frac{\delta}{\sigma/\sqrt{n}}\right)$$

$$\iff \qquad \delta \;=\; \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}(1-\alpha)$$

Therefore, with a confidence level of $1-\alpha$:

$$\mu \;\leq\; \overline{X} \;+\; \frac{\sigma}{\sqrt{n}} \cdot \Phi^{-1}(1-\alpha)$$

(5.15) Calculation of confidence intervals with MatLab.

The quantile function Φ^{-1} is named `norminv()` in MatLab. The calculation of some values of $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ and $\Phi^{-1}(1-\alpha)$ are displayed below:

```
>> alpha = [0.1, 0.05, 0.01, 0.005, 0.001]
alpha =
    0.1000    0.0500    0.0100    0.0050    0.0010
>> conf_level = (1-alpha)*100
conf_level =
   90.0000   95.0000   99.0000   99.5000   99.9000
>> norminv(1-alpha/2)
ans =
    1.6449    1.9600    2.5758    2.8070    3.2905
>> norminv(1-alpha)
ans =
    1.2816    1.6449    2.3263    2.5758    3.0902
>>
```

(5.16) Determination of confidence intervals for μ, if σ^2 is unknown.

Using the sample mean $\overline{X}$ to estimate μ and the sample variance S^2 to estimate σ^2, confidence intervals for μ can be determined by replacing the standard normal distribution used in (5.14) with the t distribution with $n-1$ degrees of freedom for the random variable:

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{(5.12)}{\sim} t_{n-1}$$

Confidence intervals for μ with a confidence level of $1-\alpha$ are:

$$\begin{aligned}
\mu &= \overline{X} \pm \frac{S}{\sqrt{n}} \cdot F_{t_{n-1}}^{-1}\left(1-\frac{\alpha}{2}\right) \\
\mu &\geq \overline{X} - \frac{S}{\sqrt{n}} \cdot F_{t_{n-1}}^{-1}\left(1-\alpha\right) \\
\mu &\leq \overline{X} + \frac{S}{\sqrt{n}} \cdot F_{t_{n-1}}^{-1}\left(1-\alpha\right)
\end{aligned}$$

(5.17) Calculation of confidence intervals with MatLab.

The calculation of some values of $F_{t_{n-1}}^{-1}\left(1-\frac{\alpha}{2}\right)$ and $F_{t_{n-1}}^{-1}\left(1-\alpha\right)$ are displayed below for $n = 3, 5$ and 10:

```
>> alpha = [0.1, 0.05, 0.01, 0.005, 0.001]
alpha =
    0.1000    0.0500    0.0100    0.0050    0.0010
>> conf_level = (1-alpha)*100
conf_level =
   90.0000   95.0000   99.0000   99.5000   99.9000
>> [tinv(1-alpha/2,2); tinv(1-alpha/2,4); tinv(1-alpha/2,9)]
ans =
    2.9200    4.3027    9.9248   14.0890   31.5991
    2.1318    2.7764    4.6041    5.5976    8.6103
    1.8331    2.2622    3.2498    3.6897    4.7809
>> [tinv(1-alpha,2); tinv(1-alpha,4); tinv(1-alpha,9)]
ans =
    1.8856    2.9200    6.9646    9.9248   22.3271
    1.5332    2.1318    3.7469    4.6041    7.1732
    1.3830    1.8331    2.8214    3.2498    4.2968
>>
```

5.1.2 Confidence intervals for σ^2

(5.18) Determination of confidence intervals for σ^2.

Using the sample variance S^2 to estimate σ^2, confidence intervals for σ^2 can be determined from the random variable:

$$Y_{n-1} \;:=\; \frac{n-1}{\sigma^2}\,S^2 \;\overset{(5.11)}{\sim}\; \chi^2_{n-1}$$

(i) Two-sided confidence intervals

Given $\alpha \in (0,1)$, put:

$$b_1 \;:= F^{-1}_{Y_{n-1}}\left(\frac{\alpha}{2}\right) \qquad \text{and} \qquad b_2 \;:= F^{-1}_{Y_{n-1}}\left(1-\frac{\alpha}{2}\right)$$

Then

$$\Pr\left(\frac{n-1}{b_2}\,S^2 \;\le\; \sigma^2 \;\le\; \frac{n-1}{b_1}\,S^2\right) \;=\; \Pr\left(b_1 \;\le\; \frac{n-1}{\sigma^2}\,S^2 \;\le\; b_2\right) \;=\; 1-\alpha$$

and we have:

$$\sigma^2 \;\in\; \left[\frac{n-1}{b_2}\,S^2,\; \frac{n-1}{b_1}\,S^2\right] \qquad \text{with confidence level } 1-\alpha$$

(ii) One-sided lower confidence intervals

Given $\alpha \in (0,1)$, put:

$$b \;:= F^{-1}_{Y_{n-1}}(\alpha)$$

Then

$$\Pr\left(\sigma^2 \;\le\; \frac{n-1}{b}\,S^2\right) \;=\; \Pr\left(\frac{n-1}{\sigma^2}\,S^2 \;\ge\; b\right) \;=\; 1-\alpha$$

and:

$$\sigma^2 \;\le\; \frac{n-1}{b}\,S^2 \qquad \text{with confidence level } 1-\alpha$$

(iii) One-sided upper confidence intervals

Given $\alpha \in (0,1)$, put:

$$b \;:= F^{-1}_{Y_{n-1}}(1-\alpha)$$

Then

$$\Pr\left(\sigma^2 \;\ge\; \frac{n-1}{b}\,S^2\right) \;=\; \Pr\left(\frac{n-1}{\sigma^2}\,S^2 \;\le\; b\right) \;=\; 1-\alpha$$

and:

$$\sigma^2 \;\ge\; \frac{n-1}{b}\,S^2 \qquad \text{with confidence level } 1-\alpha$$

5.2 Random samples from a Bernoulli distribution

(5.19) Remark.

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim \text{Bernoulli}(p)$. Then $n\overline{X} \sim \text{binomial}(n, p)$ and for sufficiently large integers n, we get:

$$\frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \;\overset{\approx}{\sim}\; \mathcal{N}(0,1)$$

(5.20) Approximate confidence intervals for p.

Using the sample mean to estimate p, approximate confidence intervals for p of confidence level $\approx (1-\alpha)$ are determined below.

(i) Two-sided confidence intervals

$$\begin{aligned}
1-\alpha &= \Pr\Big(p \in [\overline{X}-\delta, \overline{X}+\delta]\Big) = \Pr\Big(|\overline{X}-p| \le \delta\Big) \\
&= \Pr\Big(-\delta \le \overline{X}-p \le \delta\Big) \\
&= \Pr\left(-\frac{\delta}{\sqrt{p(1-p)/n}} \le \frac{\overline{X}-p}{\sqrt{p(1-p)/n}} \le \frac{\delta}{\sqrt{p(1-p)/n}}\right) \\
&\approx 2 \cdot \Phi\left(\frac{\delta}{\sqrt{p(1-p)/n}}\right) - 1
\end{aligned}$$

$$\iff \qquad \delta \;\approx\; \frac{\sqrt{p(1-p)} \cdot \Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{\sqrt{n}} \;\le\; \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{2\sqrt{n}}$$

(ii) One-sided lower confidence intervals

$$\begin{aligned}
1-\alpha &= \Pr\Big(p \in (-\infty, \overline{X}+\delta]\Big) = \Pr\Big(\overline{X}-p \ge -\delta\Big) \\
&= \Pr\left(\frac{\overline{X}-p}{\sqrt{p(1-p)/n}} \ge -\frac{\delta}{\sqrt{p(1-p)/n}}\right) \\
&\approx \Phi\left(\frac{\delta}{\sqrt{p(1-p)/n}}\right)
\end{aligned}$$

$$\iff \qquad \delta \;\approx\; \frac{\sqrt{p(1-p)} \cdot \Phi^{-1}\left(1-\alpha\right)}{\sqrt{n}} \;\le\; \frac{\Phi^{-1}\left(1-\alpha\right)}{2\sqrt{n}}$$

(iii) One-sided upper confidence intervals

$$\begin{aligned} 1-\alpha &= \Pr\Big(p \in [\overline{X}-\delta,\infty)\Big) = \Pr\Big(\overline{X}-p \le \delta\Big) \\ &= \Pr\left(\frac{\overline{X}-p}{\sqrt{p(1-p)/n}} \le \frac{\delta}{\sqrt{p(1-p)/n}}\right) \\ &\approx \Phi\left(\frac{\delta}{\sqrt{p(1-p)/n}}\right) \end{aligned}$$

$$\iff \qquad \delta \approx \frac{\sqrt{p(1-p)}\cdot\Phi^{-1}(1-\alpha)}{\sqrt{n}} \le \frac{\Phi^{-1}(1-\alpha)}{2\sqrt{n}}$$

(5.21) Approximation of n for confidence intervals of given width.

Given α, δ and an estimation $\overline{x} \approx p$ (e.g. from a small preliminary sample), an estimation for the sample size n, such that a two-sided confidence interval for p of confidence level $1-\alpha$ has width 2δ, is given by:

$$n \approx \frac{p(1-p)\cdot\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2}{\delta^2} \approx \frac{\overline{x}(1-\overline{x})\cdot\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2}{\delta^2}$$

As $p(1-p) \le \frac{1}{4}$ for all $p \in \mathbb{R}$, the estimation

$$n \lessapprox \frac{\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2}{4\delta^2}$$

holds true for all values of p.

(5.22) Example. An application of (5.21) can be used to obtain an estimation for the number n of times a fair coin must be thrown in order to see "head" in 49%–51% of all cases with a probability of at least 98%:

```
>> alpha = 0.02;
>> delta = 0.01;
>> n = norminv(1-alpha/2)^2/(4*delta^2)
n =
   1.3530e+04
```

5.3 Random samples from an exponential distribution

(5.23) Lemma. Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim \exp(\lambda)$. Then:

$$Y_{2n} \;:=\; 2\lambda \cdot \sum_{i=1}^{n} X_i \;=\; \chi^2_{2n}$$

(5.24) Confidence intervals for λ and $1/\lambda$.

Using the sample mean $\overline{X}$ to estimate $1/\lambda$, confidence intervals for $1/\lambda$ (and λ) of confidence level $1-\alpha$ are determined below.

(i) Two-sided confidence intervals

Given $\alpha \in (0,1)$, put:

$$b_1 := F^{-1}_{Y_{2n}}\left(\frac{\alpha}{2}\right) \qquad \text{and} \qquad b_2 := F^{-1}_{Y_{2n}}\left(1-\frac{\alpha}{2}\right)$$

Then

$$\Pr\left(\frac{2n\,\overline{X}}{b_2} \le 1/\lambda \le \frac{2n\,\overline{X}}{b_1}\right) = \Pr\left(\frac{b_1}{2n\,\overline{X}} \le \lambda \le \frac{b_2}{2n\,\overline{X}}\right)$$
$$= \Pr\left(b_1 \le 2\lambda n\,\overline{X} \le b_2\right) = 1-\alpha$$

and therefore:

$$1/\lambda \in \left[\frac{2n}{b_2}\cdot\overline{X},\ \frac{2n}{b_1}\cdot\overline{X}\right] \qquad \text{with confidence level } 1-\alpha$$

$$\lambda \in \left[\frac{b_1}{2n}\cdot\frac{1}{\overline{X}},\ \frac{b_2}{2n}\cdot\frac{1}{\overline{X}}\right] \qquad \text{with confidence level } 1-\alpha$$

(ii) One-sided lower confidence intervals for $1/\lambda$

Given $\alpha \in (0,1)$, put:

$$b := F^{-1}_{Y_{2n}}(\alpha)$$

Then

$$\Pr\left(1/\lambda \le \frac{2n\,\overline{X}}{b}\right) = \Pr\left(\lambda \ge \frac{b}{2n\,\overline{X}}\right) = \Pr\left(2\lambda n\,\overline{X} \ge b\right) = 1-\alpha$$

and:

$$1/\lambda \le \frac{2n}{b}\cdot\overline{X} \qquad \text{with confidence level } 1-\alpha$$

$$\lambda \ge \frac{b}{2n}\cdot\frac{1}{\overline{X}} \qquad \text{with confidence level } 1-\alpha$$

(iii) One-sided upper confidence intervals for $1/\lambda$

Given $\alpha \in (0,1)$, put:

$$b := F_{Y_{2n}}^{-1}(1-\alpha)$$

Then

$$\Pr\left(1/\lambda \geq \frac{2n\overline{X}}{b}\right) = \Pr\left(\lambda \leq \frac{b}{2n\overline{X}}\right) = \Pr\left(2\lambda n\overline{X} \leq b\right) = 1-\alpha$$

and:

$$1/\lambda \geq \frac{2n}{b} \cdot \overline{X} \qquad \text{with confidence level } 1-\alpha$$

$$\lambda \leq \frac{b}{2n} \cdot \frac{1}{\overline{X}} \qquad \text{with confidence level } 1-\alpha$$