Probability and Statistics

5 – Statistics

Stefan Heiss

Technische Hochschule Ostwestfalen-Lippe
Dep. of Electrical Engineering and Computer Science

January 02, 2024

Random Samples

Definition (5.1)

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

Random Samples

Remark (5.2)

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.

(Ω, X)

$\tilde{\Omega} := \Omega \times \Omega \times \Omega \times \cdots \times \Omega$

$\omega = (\omega_1, \omega_2, \omega_3, \cdots, \omega_n) \in \tilde{\Omega}$

$X_i(\omega) = X(\omega_i) =: x_i$ $\rightarrow$ data set: $(x_1, x_2, \ldots, x_n)$

Statistics

Definition (5.3)

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).

$$\tilde{\Omega} = \Omega^n$$

$$X_1, \quad X_2, \quad \dots, \quad X_n$$

$$\mathbb{R} \quad \mathbb{R} \quad \mathbb{R}$$

$(x_1, x_2, \dots, x_n) \leftarrow$ data set

$$\Big\downarrow S$$

$$S(X_1, \dots, X_n) : \Omega^n \to \mathbb{R}$$

Statistical Inference

Remark (5.4)

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.

Estimators

Definition (5.5)

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*.

Sample Mean

$E(X_i) = \mu$

$Var(X_i) = \sigma^2$

$Var(\overline{X}) = \frac{1}{n^2} n \cdot \sigma^2 = \frac{\sigma^2}{n}$

Definition (5.6)

The *sample mean* of a random sample $X_1, \dots, X_n$ is defined to be:

$$\overline{X} := \frac{X_1 + \cdots + X_n}{n} \xrightarrow[n \to \infty]{CLT} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Expectation and Variance of Sample Mean

Lemma (5.7)

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

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

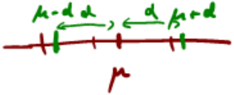
(ii) $Var(\overline{X}) = \dfrac{\sigma^2}{n}$ ✓

Central Limit Theorem

Lemma (5.8)

For a random sample with $\mu = E(X_1)$, $\sigma^2 = Var(X_1)$ and sufficiently large n, the central limit theorem 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 of Lemma (5.8)

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

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

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

Sample Variances and Sample Deviations

Definition (5.9)

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.

Sample Variances

Lemma (5.10)

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

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

Proof of Lemma (5.10)

$$
\begin{aligned}
E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) &= E\left(\sum_{i=1}^{n}(X_i^2-2\overline{X}X_i+\overline{X}^2)\right) = E\left(\sum_{i=1}^{n}X_i^2-2\overline{X}\sum_{i=1}^{n}X_i+n\overline{X}^2\right)\\
&= E\left(\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right) \quad \checkmark\\
&= \underline{nE\left(X_1^2\right)}\ \underline{-}\ \underline{nE\left(\overline{X}^2\right)} \quad \checkmark\\
&= n\left(\underline{Var(X_1)+E(X_1)^2}\ \underline{-\ Var(\overline{X})}\ \underline{-\ E(\overline{X})^2}\right)\\
&= n\left(\sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2\right) = (n-1)\,\sigma^2
\end{aligned}
$$

Random Samples with Normal Distributions

Theorem (5.11)

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:

(without proof)

$$\overline{X} \sim \mathcal{N}(\mu, \sigma/\sqrt{n}) \quad \checkmark \qquad \text{and} \qquad \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

Partial Proof of Theorem (5.11)

$$\begin{aligned}
\frac{n-1}{\sigma^2} S^2 &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2\right) \\
&= \frac{1}{\sigma^2}\left(\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) - n(\overline{X} - \mu)^2\right) \\
&= \left(\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2\right) - \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2
\end{aligned}$$

$$-2\sum_{i=1}^{n}(X_i - \mu)\cdot(\bar{X} - \mu)$$
$$= -2(\bar{X} - \mu)\cdot\sum_{i=1}^{n}(X_i - \mu)$$
$$= -2(\bar{X} - \mu)(n\cdot\bar{X} - n\mu)$$
$$= -2n(\bar{X} - \mu)^2$$

Hence

$$\frac{n-1}{\sigma^2} S^2 + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 = \left(\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2\right)$$

Partial Proof of Theorem (5.11)

without proof S^2 and $\bar{X}$ are independent

$$\frac{n-1}{\sigma^2}S^2 + \underbrace{\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2}_{\sim \chi^2_1} = \underbrace{\left(\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right)}_{\sim \chi^2_n}$$

m.g.f.:

$$\phi_1(t) \cdot \left(\frac{1}{1-2t}\right)^{1/2} = \left(\frac{1}{1-2t}\right)^{n/2} \quad , t < \frac{1}{2}$$

see (4.25) and (4.78)

$$\phi_1(t) = \left(\frac{1}{1-2t}\right)^{\frac{n-1}{2}}$$

m.g.f. for χ^2_{n-1} distribution

$$\Rightarrow \quad \frac{n-1}{\sigma^2}\cdot S^2 \sim \chi^2_{n-1}$$

Random Samples with Normal Distributions

Corollary (5.12)

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

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

Proof:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overbrace{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}^{\sim \mathcal{N}(0,1)}}{\sqrt{\underbrace{\frac{n-1}{\sigma^2} S^2}_{\sim \chi^2_{n-1}} /(n-1)}} \overset{(5.11)}{\sim} t_{n-1}$$

Confidence Intervals

Notation (5.13)

Let $\overline{x}$ denote the value of the sample mean of a random sample. Given an $\alpha \in (0,1)$ and intervals $I = I(\overline{x})$ for all such values $\overline{x}$, such that

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

then 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:

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