Statistics

From probability to statistics

In probability:

$$X_i \sim F(x) = \mathbb{P}(X \leqslant x)$$

In statistics:

$$\{X_1, \ldots, X_n\} + \text{hypothesis} \to \text{information}$$

What kind of information?

- Distribution
- Distribution property
- Functional
- Hypothesis correctness

What we will do in this course

- Estimation
  - Point estimates = best guess
  - Bayesian estimate = probabilistic guess
- Confidence sets (another way to provide probabilistic guess)
- Hypothesis testing (confidence sets inversed)

Parametric vs non-parametric statistics

- In probability, we studied different distributions, that were parametrized by one or more parameters.
- In statistics, we can think of these distributions as of **models** of the data, and try to find the **parameters** for these models best describing the data.
- Example: you own a chain of restaurants, and due to budget cuts, you would like to close one restaurant that is the least popular. In order to do that, for each restaurant you collect the data of visitor counts per day. Then you can fit a distribution to find the actual rates of visitors per day and then proceed to close the restaurant with the lowest rate.
- This is called parametric statistics.

Parametric vs non-parametric statistics

- We can also go without any model assumption. How?
- This will be non-parametric statistics.

Estimation

Estimation

Consider sample $X_1, X_2, \ldots, X_n$ i.i.d. (simple sample). Consider its true distribution $F_\theta(x)$ (unknown) parametrized by θ. In case of parametric statistics this will be the set of distribution parameters, in case of non-parametric statistics it will be the distribution itself (next time).

Consider an estimate $\hat{\theta}$. It is a **random variable**, since it is a function of sample, which is random. How **good** is this estimate?

Bias

Bias is one measure of goodness of fit, describing the systematic error (not related to randomness).

$$\text{bias}\left(\hat{\theta}\right) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

An estimate is called **unbiased** if bias is zero:

$$\mathbb{E}\left[\hat{\theta}\right] = \theta$$

Unbiasedness is a desirable quality, however it is often omitted to reduce the other measure of goodness of fit.

Example 1

Example with restaurant. Consider simple sample $X_1, \ldots, X_n \sim Pois(\lambda)$. We would like to estimate the parameter of Poisson distribution and we chose the following estimator:

$$\widehat{\lambda} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

What would be its bias?

Solution 1

$$\mathbb{E}\left[\widehat{\lambda}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[X_k] = \frac{1}{n}\sum_{k=1}^{n}\lambda = \lambda$$
$$\text{bias}\left(\widehat{\lambda}\right) = \mathbb{E}\left[\widehat{\lambda}\right] - \lambda = 0$$

Standard error

Bias describes the tendency of our estimate, but what about its noisiness? This is where standard error comes in handy.

$$\text{se}\left(\hat{\theta}\right) = \sqrt{\mathbb{V}\text{ar}\left(\hat{\theta}\right)}$$

Standard error is very important. We can have a perfectly unbiased estimator with large standard error.

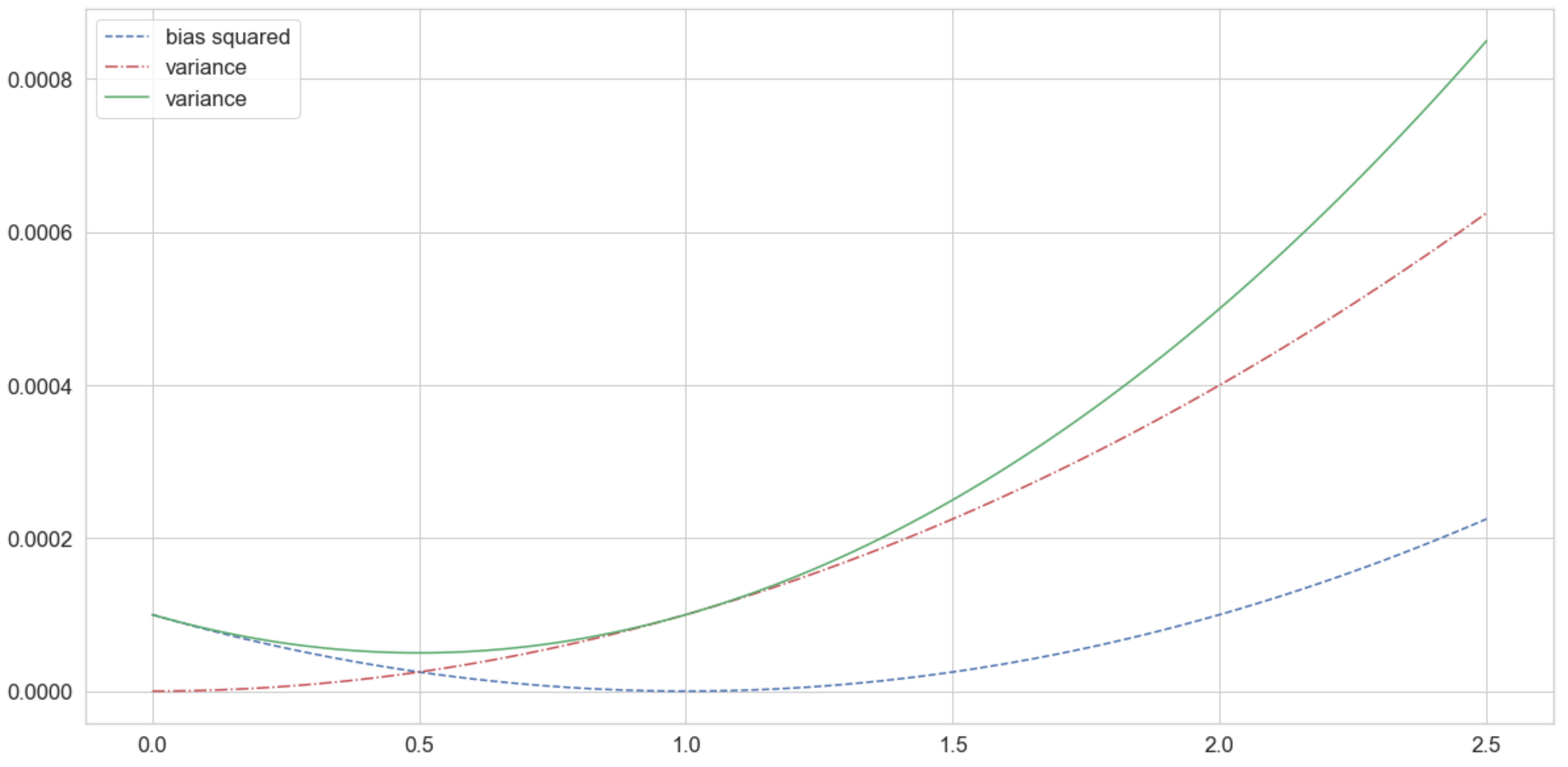
In [3]:
```
import numpy as np
import scipy.stats as sts
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid", font_scale=1.5)
sns.despine()

%matplotlib inline
```

In [21]:
```
lbd = 1/100
n = 100
x = sts.poisson(lbd).rvs(n)
lbd_hat = np.mean(x)
scaling_factors = np.linspace(0, 2.5, 100)
biases = lbd * (scaling_factors - 1)
ses = scaling_factors * np.sqrt(lbd / n)
fig, ax = plt.subplots(figsize=(20,10))
ax.plot(scaling_factors, biases ** 2, "b--", label="bias squared")
ax.plot(scaling_factors, ses ** 2, "r-.", label="variance")
ax.plot(scaling_factors, biases ** 2 + ses ** 2, 'g-', label="variance")
ax.legend()
```

Out[21]:
```
<matplotlib.legend.Legend at 0x13b342640>
```



Bias-variance tradeoff

Actually bias and standard error are strongly connected:

$$\text{MSE}\left(\hat{\theta}\right) = \text{bias}\left(\hat{\theta}\right) + \text{se}^2\left(\hat{\theta}\right)$$

Example 2

Example with restaurant. Consider simple sample $X_1, \ldots, X_n \sim Pois(\lambda)$. We would like to estimate the parameter of Poisson distribution and we chose the following estimator:

$$\widehat{\lambda} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

What would be its standard error?

Solution 2

$$\mathbb{V}\text{ar}\left(\widehat{\lambda}\right) = \mathbb{E}\left[\widehat{\lambda}^2\right] - \left(\mathbb{E}\left[\widehat{\lambda}\right]\right)^2$$
$$\mathbb{E}\left[\widehat{\lambda}^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^2\right] = \mathbb{E}\left[\frac{1}{n^2}\sum_{k,m=1}^{n} X_k X_m\right] =$$
$$= \frac{1}{n^2}\left(\sum_{k=1}^{n}\mathbb{E}\left[X_k^2\right] + \sum_{k\neq m} E[X_k]E[X_m]\right) =$$
$$= \frac{1}{n^2}\left(n\cdot(\lambda^2+\lambda) + n(n-1)\cdot\lambda^2\right) = \frac{1}{n}\left(n\lambda^2+\lambda\right)$$
$$\text{se}\left(\widehat{\lambda}\right) = \sqrt{\mathbb{V}\text{ar}\left(\widehat{\lambda}\right)} = \sqrt{\frac{1}{n}\left(n\lambda^2+\lambda\right) - \lambda^2} = \sqrt{\lambda/n}$$

Consistency and strong consistency

An estiamtor $\hat{\theta}$ is called consistent if

$$\hat{\theta} \xrightarrow{P} \theta$$

What is the base of the limit? What type of convergence is it?

An estiamtor $\hat{\theta}$ is called strongly consistent if

$$\hat{\theta} \xrightarrow{a.s.} \theta$$

What type of convergence is it?

Example 3

Example with restaurant. Consider simple sample $X_1, \ldots, X_n \sim Pois(\lambda)$. We would like to estimate the parameter of Poisson distribution and we chose the following estimator:

$$\widehat{\lambda} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Is this estimator consistent?

Solution 3

As I hope everyone remembers, convergence in probability is that convergence in probability is the weakest of all, and follows from other types of convergence.

We have $\text{bias}\left(\widehat{\lambda}\right) = 0$ and $\text{se}\left(\widehat{\lambda}\right) = \sqrt{\lambda/n}$. From bias-variance tradeoff, we can see that

$$\text{MSE}\left(\widehat{\lambda}\right) = \text{bias}^2\left(\widehat{\lambda}\right) + \text{se}^2\left(\widehat{\lambda}\right) = \lambda/n$$

We can see that $\text{MSE}\left(\widehat{\lambda}\right) \to 0$, as $n \to \infty$. This means convergence in mean-squared, and convergence in probability follows from it unconditionally.

Strong consistency follows from strong LLN.