Applied Statistics

Sofia Triantafillou

sof.triantafillou@gmail.com

University of Crete
Department of Mathematics and Applied Mathematics

Lecture Summary

- The sample mean (Chapter 6.2 - Properties of the sample mean).
- Central Limit Theorem (Chapter 6.3).
- Statistical Inference (Chapter 7.1).
- The likelihood function (part of Chapter 7.2).

Recap: Normal Distribution

- Standard normal: $\mathcal{N}(0,1) : f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
- Normal with mean μ and variance σ^2:
  $\mathcal{N}(\mu,\sigma^2) : f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
- $E(X) = \mu$, $Var(X) = \sigma^2$.

Theorem (Linear transformations of a normal are normal)

If $X \sim \mathcal{N}(\mu,\sigma^2)$, *then* $\alpha X + \beta \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$

Theorem (The sum of independent normals is normal)

If the random variables $X_1, \ldots, X_k$ *are independent and if* $X_i \sim N(\mu_i, \sigma_i^2)$ *then* $X_1 + \cdots + X_k \sim N(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2)$

Sample Mean

Definition (Sample mean)

Let $X_1, \ldots X_n$ be random variables. Their average

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

is called their *sample mean*.

- Find $E(\overline{X}_n)$ and $Var(\overline{X}_n)$ when X_i are independent and follow the same distribution.

Properties of the sample mean

Theorem (Mean and variance of the sample mean)

Let $X_1, \ldots X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then $E(\overline{X}_n) = \mu$, and $Var(\overline{X}_n) = \sigma^2/n$.*

*also known as: $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d).

Central Limit Theorem

- $S_n = \sum_{i=1}^{n} X_i$, mean $n\mu$, variance $n\sigma^2$.
- $\bar{X}_n = \frac{S_n}{n}$, mean μ variance $\frac{\sigma^2}{n}$.
- $\frac{S_n}{\sqrt{n}}$, mean $\mu\sqrt{n}$, variance σ^2.
- $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, mean 0, variance 1.

Central Limit Theorem

Theorem (Central Limit Theorem)

If the random variables $X_1, \ldots, X_n$ form a random sample of size n from a given distribution with mean μ and variance σ^2 $(0 < \sigma^2 < \infty)$, then for each fixed number x

$$lim_{n\to\infty} P(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq x) = \Phi(x),$$

where Φ denotes the c.d.f. of the standard normal distribution.

OR

$S_n = X_1 + \ldots X_n, \qquad Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}, \qquad Z \sim \mathcal{N}(0,1).$

For every z,

$$lim_{n\to\infty} P(Z_n \leq z) = P(Z \leq z)$$

Central Limit Theorem

- is pretty great! Holds for ANY type of distribution of X_i with finite variance, as long as X_i are i.i.d.
  - Results for non i.i.d. under stronger assumptions.
- Convergence to CDF, not PDF.
  - In practice, we use pdf for illustration.
- Proof requires moment generating functions.
- Is a good approximation even for small n in practice (see recitation tomorrow).
- Has many practical uses.

Central Limit Theorem: Approximation for the Binomial.

- $X_1, \ldots, X_n \sim Bernoulli(p)$.
- $S_n \sim Binom(n, p)$
- $\frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{S_n - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1)$
- Example: Assume $X_i \sim Bernoulli(0.5)$, $n = 36$. Find $P(\text{"no more than 21 successes"})$

Correction for continuity:

If a continuous variable Y with pdf g provides a good approximation for the discrete variable X, with pmf f, then

$$P(a \leq X \leq b) = P(a - 1/2 \leq Y \leq b + 1/2)$$

CDF of the normal

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,9	0,8159	0,8185	0,8212	0,8238	0,8263	0,8289	0,8314	0,8339	0,8364	0,8389
1	0,8413	0,8437	0,8461	0,8484	0,8508	0,8531	0,8554	0,8576	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8707	0,8728	0,8749	0,8769	0,879	0,881	0,8829
1,2	0,8849	0,8868	0,8887	0,8906	0,8925	0,8943	0,8961	0,8979	0,8997	0,9014
1,3	0,9032	0,9049	0,9065	0,9082	0,9098	0,9114	0,9130	0,9146	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9250	0,9264	0,9278	0,9292	0,9305	0,9318
1,5	0,9331	0,9344	0,9357	0,9369	0,9382	0,9394	0,9406	0,9417	0,9429	0,9440
1,6	0,9452	0,9463	0,9473	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9544
1,7	0,9554	0,9563	0,9572	0,9581	0,9590	0,9599	0,9608	0,9616	0,9624	0,9632
1,8	0,9640	0,9648	0,9656	0,9663	0,9671	0,9678	0,9685	0,9692	0,9699	0,9706
1,9	0,9712	0,9719	0,9725	0,9732	0,9738	0,9744	0,975	0,9755	0,9761	0,9767
2	0,9772	0,9777	0,9783	0,9788	0,9793	0,9798	0,9803	0,9807	0,9812	0,9816
2,1	0,9821	0,9825	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9853	0,9857
2,2	0,9861	0,9864	0,9867	0,9871	0,9874	0,9877	0,9880	0,9884	0,9887	0,9889
2,3	0,9892	0,9895	0,9898	0,9901	0,9903	0,9906	0,9908	0,9911	0,9913	0,9915
2,4	0,9918	0,9920	0,9922	0,9924	0,9926	0,9928	0,9930	0,9932	0,9934	0,9936
2,5	0,9937	0,9939	0,9941	0,9943	0,9944	0,9946	0,9947	0,9949	0,9950	0,9952
2,6	0,9953	0,9954	0,9956	0,9957	0,9958	0,9959	0,9960	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9972	0,9973

Example

- You are doing a poll on "ratio of people agree with the lockdown measures".
- True ratio: p, estimate $\bar{X}_n$ after asking n people.
- You want $|\bar{X}_n - p|$ to be small (e.g., less than 1%)

Example

- You are doing a poll on "ratio of people agree with the lockdown measures".
- True ratio: p, estimate $\bar{X}_n$ after asking n people.
- You want $|\bar{X}_n - p|$ to be small (e.g., less than 1%)
- You want to estimate

$$P(|\bar{X}_n - p| \geq 0.01) \leq 0.05$$

- Apply CLT:

Recap

Probability and Random Variables

Sofar, we have seen how random variables can describe some attributes of a random experiment, and how we can use the language of probability to describe the distributions of random variables, and some of their large sample properties.

Next

We will now start talking about how we can make inferences about the distributions of random variables based on our observations (data). This is the area of statistical inference.

Statistical Inference

So far: We have seen statistical models in the form of probability distributions: $f(x|\theta)$
We use: θ for parameters, Ω for the parameter space.

Examples

- The height of a student is approximately normal with mean θ and some known variance.
- The number of people that have a disease out of a group of N people follows the Binomial (N, θ) distribution.
- More distributions tomorrow.

In practice, we do not know θ.

Statistical Inference

What can we *infer* about θ given the observed data?
Assuming that we observe random variables $X_1, \ldots, X_n$ following some distribution with parameter θ, what conclusions can we draw about parameter θ?

Example

Say I take a random sample of 100 people and test them all for a disease. If 3 of them have the disease, what can I say about $\theta =$ the prevalence of the disease in the population?

Statistical Inference

Example

Say I take a random sample of 100 people and test them all for a disease. If 3 of them have the disease, what can I say about = the prevalence of the disease in the population?

- Say I estimate θ as $3/100 = 0.3$ or 3%.
- How sure am I about this number?
- I want uncertainty bounds on my estimate.
- Am I confident that the prevalence of the disease is higher than 2% ?

Statistical Inference tasks

Prediction

Predict random variables that have not yet been observed. e.g., if we test 40 more people for the disease, how many people do we predict have the disease?

Estimation

Estimate (predict) the unknown parameter θ.
e.g., we estimated the prevalence of the disease as $\hat{\theta}$.

Statistical Inference tasks

Statistical Decision Problems

Hypothesis testing, decision theory e.g., If the disease affects 2% or more of the population, the state will launch a costly public health campaign. Can we be confident that θ is higher than 2% ?

Experimental Design

What and how much data should we collect?
e.g., how do I select people in my clinical trial? How many do I need to make a decision based on that data? Often limited by budget/ethical constraints.