476 Statistics, Spring 2022.
Lecturer: Sergey Nadtochiy.

Lecture 1. Statistical inference, distributions of popular statistics of a normal sample. (Sections 7.1–7.2)

1 Statistical inference

What is the main question of statistics? In plain English, based on the observations, we want to deduce (infer) information about how these observations were generated. But this is too general and vague. More specifically, statistics addresses the above question in the context of Probability Theory: i.e., when the observations are viewed as specific values of random variables.

In the basic case,

- we are given a sample of i.i.d. random variables $Y_1, \ldots, Y_n$,
- we make an assumption that the distribution of Y_i belongs to a parametric family of distr.'s, parameterized by θ (e.g. $\{N(\mu, 1)\}$, where $\theta = \mu \in \mathbb{R}$),
- and we would like to infer as much information as possible about the true value of θ using the sample values.

Examples of this are numerous: trend in the global temperature or in a stock price (this is estimation of the true expected value of annual change in the temperature), a proportion of voters that will vote for a given candidate (this is estimation of the probability that a randomly chosen voter chooses a specific candidate), is it true that a change in the company's website increases traffic to this website? (this is testing a hypothesis that the expected number of daily visits is higher after the change than before the change was implemented), etc.

There are many different ways to estimate unknown parameters and test hypotheses. In fact, one may come up with pretty much any estimation algorithm. However, not all of the algorithms are equally good. **The main point of this course is to identify the desirable properties of the estimators and tests and to design the estimators and tests that do have these properties!** This is the subject of Theoretical (or Mathematical) Statistics. In Applied Statistics, you will see how these methods are applied to real data and how they can be combined in order to build more complex predictive models.

In order to estimate an unknown parameter θ, we construct functions of the sample $Y_1, \ldots, Y_n$.

Def 1. *Any function of $(Y_1, \ldots, Y_n)$ is called a statistic of this sample.*

2 Popular statistics

Note that any statistic is a **random variable**. In practice, we observe particular realizations of these random variables.

Not every statistic is relevant - it has to contain information about the true θ. Let us discuss examples of relevant θ and the associated statistics

- $\theta = \mathbb{E}Y_i$. Then, the sample mean $\bar{Y}$ is a relevant statistic, due to the Law of Large Numbers LLN.

  Thm 1. *(Strong LLN) Assume we are given i.i.d. r.v.'s $Y_1, \ldots, Y_n$, with finite mean μ. Then,*

$$\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

converges to μ with probability one.

- $\theta = V(Y_i)$. Then

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\mu)^2$$

  is a relevant statistic, because it converges to $V(Y_i)$ due to LLN. Indeed, the above is a sample mean of the new i.i.d. sample $\{Z_i := (Y_i-\mu)^2\}$, and $\mathbb{E}Z_i = V(Y_i)$.

- $\theta = (\theta_1, \theta_2) = (\mathbb{E}Y_i, V(Y_i))$. Then

$$\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

  are relevant statistics. The first one is relevant due to LLN. For the second one, we have to work a bit.

  Thm 2. *Assume we are given i.i.d. r.v.'s $Y_1, \ldots, Y_n$, with finite mean (expected value) μ and variance σ^2. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2 \to \sigma^2$$

  Proof:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu-(\bar{Y}_n-\mu))^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i-\mu)^2 - 2(\bar{Y}_n-\mu)(Y_i-\mu) + (\bar{Y}_n-\mu)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2 - 2(\bar{Y}_n-\mu)\left[\frac{1}{n}\sum_{i=1}^{n}Y_i-\mu\right] + (\bar{Y}_n-\mu)^2 \to \sigma^2 - 0 + 0.$$

  ∎

- $\theta = \mathbb{P}(Y_i \in A)$, for some fixed $A \subset \mathbb{R}$ (e.g. $A = \{1\}$ may denote the success of a Bernoulli trial). Then

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad Z_i := \mathbf{1}_A(Y_i),$$

  is the relevant statistic for θ. Recall that the indicator of set A at point x, $\mathbf{1}_A(x)$, takes value 1 if $x \in A$ and 0 otherwise. As the new sample $\{Z_i\}$ is i.i.d. (every Z_i is the same function of Y_i), LLN implies that $\bar{Z} \to \mathbb{E}Z_i = \mathbb{P}(Y_i \in A) = \theta$, which explains why $\bar{Z}$ is a relevant statistic.

- Minimum, maximum and median of a sample may also be relevant (we'll see this later in the course).

So far, we are being vague about how exactly a statistic is used to infer information about the unknown parameter theta (although it is clear in the simplest cases described above). But it is important to realize that any statistic is a r.v., as a function of other r.v.'s. Thus, in order to develop the desired estimates, it is often important to know the **distribution of a relevant statistic**. The methods of Probability Theory can be used to find the distribution of a function of random variables. This task is not easy in general, but it simplifies significantly if we deal with i.i.d. normal random variables.

3 Distributions of popular statistics of a normal sample

Recall the definition of the normal distribution with mean μ and variance σ^2, denoted $N(\mu, \sigma^2)$.

Thm 3. *Assume $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$. Then,*

$$\bar{Y} \sim N(\mu, \sigma^2/n).$$

Proof:

$$\mathbb{E}\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}Y_i = \mu,$$

$$V(\bar{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sigma^2 n = \sigma^2/n.$$

Note that the above is true for any i.i.d. sample $\{Y_i\}$, even if it is not normal.

For a normal sample, we have, in addition that $\bar{Y}$ is normal. To see this, we recall the fact from Probability Theory that a sum of independent normals is normal. This fact follows easily from the Moment Generating Function (MGF). Assume that $Y_i \sim N(\mu_i, \sigma_i^2)$. Then, the MGF of $Y := Y_1 + \cdots + Y_n$ is

$$m(t) = \mathbb{E}e^{t(Y_1+\cdots+Y_n)} = \prod_{i=1}^{n} \mathbb{E}e^{tY_i} = \prod_{i=1}^{n} e^{t\mu_i + t^2\sigma_i^2/2}$$

$$= e^{t\sum_{i=1}^{n}\mu_i + t^2(\sum_{i=1}^{n}\sigma_i^2)/2} = e^{t\mu + t^2\sigma^2/2},$$

where $\mu := \sum_{i=1}^{n}\mu_i$ and $\sigma^2 := \sum_{i=1}^{n}\sigma_i^2$ are the mean and variance of Y. The above is a normal MGF, hence Y is normal.

Using the above fact, we conclude that $\sum_{i=1}^{n} Y_i$ is normal, and hence $\bar{Y}$ is norma (as it is given by $\sum_{i=1}^{n} Y_i$ multiplied by a constant). ∎

It follows that

$$\sqrt{n}\frac{\bar{Y} - \mu}{\sigma} \sim N(0, 1).$$

Ex 1. *Consider a factory that fills soda into bottles. Bottling machine fills Y_i ounce in the i-th bottle, $\{Y_i\}$ are i.i.d., $Y_i \sim N(\mu, 1)$, where μ is unknown. Consider a sample of size 9 and the associated $\bar{Y}$.*

Q 1. *What is the prob. that $\bar{Y}$ is within 0.3 ounce of μ?*

Using the above theorem, we obtain

$$\mathbb{P}(|\bar{Y} - \mu| \leq 0.3) = \mathbb{P}(-0.9 \leq \frac{\bar{Y} - \mu}{1/3} \leq 0.9) = \Phi(0.9) - \Phi(-0.9) = 2\Phi(0.9) - 1 \approx 0.6318,$$

where Φ is a standard normal cdf, and we recalled the following useful symmetry of norma cdf:

$$\Phi(-x) = 1 - \Phi(x), \quad \forall\, x \in \mathbb{R}.$$

Q 2. *How large does the sample have to be to obtain an estimate that is within 0.3 of μ with prob. at least 0.95?*

$$\mathbb{P}(|\bar{Y}_n - \mu| \le 0.3) \approx 0.95,$$
$$\mathbb{P}(|\bar{Y}_n - \mu|/(1/\sqrt{n}) \le 0.3\sqrt{n}) = 2\Phi(0.3\sqrt{n}) - 1 \approx 0.95,$$
$$\Phi(0.3\sqrt{n}) \approx 1.95/2 = 0.975.$$

The equation $\Phi(z) = 0.975$ is solved by z equal to the 0.975-quantile of the standard normal distribution, which is approximately 1.96. Therefore,

$$0.3\sqrt{n} \approx 1.96, \quad n = 43.$$

3.1 Statistics with χ^2 distribution

As before, we assume everywhere in this subsection that $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$.

Recall the definition of χ^2 distribution.

Def 2. *If $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$, then $X := X_1^2 + \cdots X_n^2$ has χ^2 distribution with n degrees of freedom, denoted $X \sim \chi^2(n)$.*

Recall that the statistic

$$U := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$$

approximates σ^2 as $n \to \infty$. Notice also that

$$\frac{n}{\sigma^2}U = \sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(n),$$

because $\{\frac{Y_i-\mu}{\sigma}\}$ are i.i.d. $N(0,1)$. This provides one explanation for why χ^2 appears naturally in statistics.

Let us recall the properties of χ^2.

- $\chi^2(n) = Gamma(n/2, 2)$.
- For $X \sim \chi^2(n)$, we have:

$$\mathbb{E}X = n, \quad V(X) = 2n.$$

  Exercise 1. *Derive the above formulas for $\mathbb{E}X$ and $V(X)$.*

- The pdf of $X \sim \chi^2(n)$ is

$$f(x) = e^{-x/2}x^{n/2-1}\frac{1}{2^{n/2}\Gamma(n/2)}, \quad x \ge 0,$$

  and $f(x) = 0$ for $x < 0$, where Γ is the Gamma-function.

We can find the quantiles of χ^2 distribution from the book.

Recall that often we don't know the true mean μ. Then, the relevant statistic for the variance σ^2 is the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

It turns out that the normalized sample variance

$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2$$

also has χ^2 distribution. However, because $\{\frac{Y_i-\bar{Y}}{\sigma}\}$ are not independent (and their variance is not exactly one), the poof of this fact is not straightforward and the associated χ^2 distribution has $n-1$ degrees of freedom.

Thm 4. *Assume $Y_1,\dots,Y_n$ are i.i.d. $N(\mu,\sigma^2)$. Then,*

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2(n-1).$$

In addition, $\bar{Y}$ and S^2 are independent.

The next example illustrates when the knowledge of the distribution of $\frac{(n-1)}{\sigma^2}S^2$ may become useful.

Ex 2. *Bottling machine fills Y_i ounce in the ith bottle, $\{Y_i\}$ are i.i.d., $N(\mu,1)$, where μ is unknown. Consider a sample of size 10 and the associated $\bar{Y}$.*

Q 3. *Find $b_1 < b_2$ s.t.*

$$\mathbb{P}(b_1 \le S^2 \le b_2) = 0.9.$$

$$0.9 = \mathbb{P}(b_1 \le S^2 \le b_2) = \mathbb{P}(\frac{10-1}{1^2}b_1 \le \frac{10-1}{1^2}S^2 \le \frac{10-1}{1^2}b_2) = F(9b_2) - F(9b_1),$$

where F is the cdf of $\chi^2(9)$ and we used the fact that $\frac{10-1}{1^2}S^2 = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(9)$.

Denote $z_1 := 9b_1$, $z_2 := 9b_2$ and notice that we can choose them as quantiles at levels p and $0.9+p$ respectively, for the above equation to hold. But there is still ambiguity in choosing the values of z_1, z_2. The canonical choice is such that $p = 1-(0.9+p)$, which means that the probability of S^2 being above the interval $[b_1,b_2]$ is equal to the probability of S^2 being below this interval (and each such probability must be 0.05, as their sum must be 0.1). Thus, $p = 0.05$ and $z_1 = 3.325$, $z_2 = 16.919$ (from the tables for $\chi^2(9)$ quantiles in the textbook).

Thus,

$$9b_1 = 3.325, \quad 9b_2 = 16.919,$$

$$b_1 \approx 0.369, \quad b_2 \approx 1.88.$$

3.2 A statistic with Student distribution (aka T-distribution)

As before, we assume everywhere in this subsection that $Y_1,\dots,Y_n$ are i.i.d. $N(\mu,\sigma^2)$.

Def 3. *Assume that $Z \sim N(0,1)$ and $W \sim \chi^2(n)$ are indep. Then, the random variable*

$$\frac{Z}{\sqrt{W/n}}$$

has Student distr. (T-distr.) with n degrees of freedom, denoted $T(n)$.

T-distribution shows up when we do not know neither μ nor σ^2. For example, the following statistic

$$\sqrt{n}\frac{\bar{Y}-\mu}{S},$$

with $S := \sqrt{S^2}$, often becomes relevant in such cases (see the example below). Notice that

$$\sqrt{n}\frac{\bar{Y}-\mu}{S} = \frac{\sqrt{n}(\bar{Y}-\mu)/\sigma}{\sqrt{\left(\frac{n-1}{\sigma^2}S^2\right)/(n-1)}} = \frac{Z}{\sqrt{W/(n-1)}},$$

where $Z := \sqrt{n}(\bar{Y}-\mu)/\sigma \sim N(0,1)$ and $W := \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$ (by Theorem 4), and Z, W are independent (also by Theorem 4). Then, by the above definition, we conclude that

$$\sqrt{n}\frac{\bar{Y}-\mu}{S} \sim T(n-1).$$

Here are some basic properties of T-distribution.

- If $X \sim T(n)$, then

  $$\mathbb{E}X = 0, \quad V(X) = \frac{n}{n-2}.$$

  For $n \leq 2$, the variance is infinite.

  It is easy to derive the above formula for the expectation:

  $$\mathbb{E}X = \mathbb{E}\frac{Z}{\sqrt{W/\nu}} = \mathbb{E}Z\,\mathbb{E}\frac{1}{\sqrt{W/\nu}} = 0,$$

  where $Z \sim N(0,1)$, $W \sim \chi^2(\nu)$, and we used their independence.

  Exercise 2. *Derive the above formula for $V(X)$.*

- The pdf of $T(n)$ looks similar to the normal. In particular, it is symmetric and

  $$F(-x) = 1 - F(x), \quad \forall\, x \in \mathbb{R},$$

  where F is the cdf of $T(n)$.

- As $n \to \infty$, W/n converges to 1 by LLN (where $W \sim \chi^2(n)$), hence, $T(n)$ becomes similar to $N(0,1)$.

- For the moderate values of n, the pdf of $T(n)$ has heavier tails than $N(0,1)$, which implies that the quantiles of $T(n)$ are farther away from 0 than the quantiles of $N(0,1)$ at the same levels. Quantiles and cdf values of $T(n)$ are in the book.

Ex 3. *Strength of a piece of wire is $N(\mu,\sigma^2)$, with unknown (μ,σ^2). A sample of 6 pieces is chosen at random, and the observed value of the statistic S is 1.*

Q 4. *Find the realized value of a statistic $b > 0$ s.t.*

$$\mathbb{P}(|\bar{Y}-\mu| > b) = 0.1.$$

The above is equivalent to

$$0.9 = \mathbb{P}\left(-b\frac{\sqrt{6}}{1} \le \frac{\sqrt{6}}{1}(\bar{Y} - \mu) \le b\frac{\sqrt{6}}{1}\right) = F(b\sqrt{6}) - F(-b\sqrt{6}) = 2F(b\sqrt{6}) - 1,$$

where F *is the cdf of* $T(5)$ *and we recalled that* $\frac{\sqrt{6}}{1}(\bar{Y} - \mu) = \frac{\sqrt{n}}{S}(\bar{Y} - \mu) \sim T(5)$*.*

To solve the equation

$$F(z) = 1.9/2 = 0.95,$$

we choose z *to be the* 0.95*-quantile of* $T(5)$*, which gives us* $z \approx 2$*.*

Thus, $b = 2/\sqrt{6}$ *and the error of* $\bar{Y}$ *is bounded from above by* $2/\sqrt{6}$ *with probability* 0.9*.*