Math 345 - Notes
Confidence intervals, hypothesis testing & statistical inference
September 26, 2018

Estimating parameters

Suppose we randomly sample data from a (much larger) population, and assume the sample describes the population. Conclusion about the sample will let us draw conclusions about the population, with some confidence.
The main goal in statistical inference is to gain knowledge of the data by focusing on certain parameters that describe it. These parameters need to be estimated and then the estimate is tested using confidence intervals, hypothesis testing etc. The most common parameters to be estimated are listed below:

I. Parameters indicating central tendency of data: if $x_1, x_2, \dots, x_n$ denote our data points,

   (a) The **mode** is the most commonly occurring data set in the sample.

   (b) The **median** is the data point in the middle of the ordered list of data. If the size of the data set is even, the median is the average of the two elements in the middle of the ordered set.

   (c) The **mean** is the average of the data points.

$$\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}.$$

II. Parameters describing the spread of the data:

   (a) The **range** is the difference between the largest element and the smallest element in the data set.

   (b) The **variance** of the set is the average of squared distances from the mean.

$$s_X^2 = \frac{(x_1 - \overline{x}_n)^2 + \cdots + (x_n - \overline{x}_n)^2}{n-1},$$

   where we divide by $n-1$ in order to get an unbiased estimator. Note that or large n, it should make very little difference if we divide by n instead of $n-1$.

   (c) The **standard deviation** is the square root of variance.

We can define the mean, variance and standard deviation for random variables as well.

- The mean, also called the expectation or the expected value, is defined as

$$E[X] = \sum_{\text{all } k} kP(X = k) \qquad \text{for discrete random variables,}$$

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx \qquad \text{for continuous random variables with density function } f(x).$$

- The variance is given by

$$Var(X) = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2.$$

More precisely,

$$\begin{array}{lll} Var(X) & = \left[\sum k^2 P(X=k)\right] - \left[\sum kP(X=k)\right]^2 \ , & \text{for discrete random variables,} \\ Var(X) & = \left[\int_{-\infty}^{\infty} x^2 f(x)\, dx\right] - \left[\int_{-\infty}^{\infty} x f(x)\, dx\right]^2 \ , & \text{for continuous random variables.} \end{array}$$

Suppose we sample a quantity that is distributed according to an unknown probability distribution function. Sampling allows us to estimate parameters of the unknown distribution, such as mean and standard deviation. Furthermore, in what we will see below, the *average* of the quantity we are interested in describing will be distributed according to a normal random variable. That means the estimated parameters will be sufficient to approximate the average of the data points.

Central Limit Theorem (CLT)

The central limit theorem states that given a large set of data, its average will be distributed according to a bell curve, irrelevant of what kind of data it is. For example, we can look at values of a roll of a 6-six sided die, or at how often a six comes up, or how long it takes to hit the target at darts, this data will be distributed according to a normal distribution, once we repeat the experiment enough times. The only difference is that the bell curves might have different mean and standard deviation. More formally, CLT says

Theorem (Central Limit Theorem): If $X_1, X_2, \ldots$ are independent trials of an experiment, each trial having the same distribution with expectation μ (finite) and variance σ^2 with $0 < \sigma^2 < \infty$, and we let

$$S_n = X_1 + X_2 + \cdots + X_n$$

and $z = N(0,1)$, then $S_n \approx N(n\mu, n\sigma^2)$ and

$$P\left(a \le \frac{S_n - n\mu}{\sqrt{n}\,\sigma} \le b\right) \approx P(a \le z \le b), \qquad \text{for } n \text{ very large.}$$

Example: Roll 10 fair 6-sided dice. Approximate the probability that the sum is between 30 and 40. Let $X_1, X_2, \ldots X_{10}$ be the value on die $1, 2, \ldots, 10$. Then X_k have the same mean and variance,

$$E[X_1] = 1 \cdot P(X_1 = 1) + 2 \cdot P(X_1 = 2) + \cdots + 6 \cdot P(X_1 = 6) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

$$\begin{aligned} Var(X_1) &= (1-3.5)^2 \cdot P(X_1 = 1) + (2-3.5)^2 \cdot P(X_1 = 2) + \cdots + (6-3.5)^2 \cdot P(X_1 = 6) \\ &= (1-3.5)^2 \cdot \frac{1}{6} + (2-3.5)^2 \cdot \frac{1}{6} + \cdots + (6-3.5)^2 \cdot \frac{1}{6} = \frac{35}{12} \end{aligned}$$

Therefore, by CLT, $S_{10} \approx N(35, 350/12)$ and using $z = N(0,1)$,

$$P\left(30 \le S_{10} \le 40\right) = P\left(\frac{30-35}{\sqrt{\frac{350}{12}}} \le \frac{S_{10}-35}{\sqrt{\frac{350}{12}}} \le \frac{40-35}{\sqrt{\frac{350}{12}}}\right) = P(-.925 < z < .925) = .644$$

Hypothesis Testing

Suppose our goal in estimating a quantity is to verify a hypothesis such as: the die if fair; the average value a 6-sided die lands on is 4, etc. Then we settle on which parameter to measure, sample the parameter and decide if the hypothesis is false, based on our sampled data. The parameter that we measure is called a **sample statistic**. We call our hypothesis **the null hypothesis** and denote it by H_0. We test it against an **alternative hypothesis**, which we denote by H_1. Our goal is to compute probabilities that lead us to **reject** or **not reject** the null hypothesis in favor of the alternative one.

$$\begin{aligned} H_0 &: \text{null hypothesis} \\ H_1 &: \text{alternative hypothesis} \end{aligned}$$

Outcomes in hypothesis testing:

- reject H_0 in favor of H_1. This means that we conclude H_0 is false.
- not reject H_0 in favor of H_1. This means means we think H_0 could still be true, but we **never** conclude H_0 is true in hypothesis testing, since there still is a small probability that H_0 is false (depends on the α-level).

Example: We want to verify if a coin is fair. Let p be the probability of the coin landing on heads. Let H_0 be the null hypothesis "$p = 0.5$" and we will test it against alternate hypothesis such as H_1 : "$p \neq 0.5$". One can test against other alternate hypotheses, such as H_2 : "$p > 0.5$" or H_3 : "$p < 0.5$."

There are two types of error one can make in hypothesis testing.

- Error of Type I, or at α-level (alpha level), measures the probability that we reject the null hypothesis H_0 even when true. For example, our data suggests that the coin is not fair, when in fact it is. Note that this can only happen if an unlikely event occurred, such as getting all heads in 100 coin flips.

  $$\alpha = \text{the maximum error of Type I that we are willing to accept in our measurement}$$

  is called the α-level, or significance level α for the test.

  Common significance levels:

  (i) $\alpha = 0.05$ corresponds to 95% confidence to make the right decision. In particular, when we test against an alternate hypothesis like H_1, which is two-tailed, we equate the 95% confidence to make the right decision with the sampled statistic falling into the 95% confidence interval. Hypotheses H_2 and H_3 are slightly different since they are one-sided, so we would not use confidence intervals for them.

  (ii) $\alpha = 0.01$ corresponds to 99% confidence to make the right decision.

- Error of Type II, or at β-level (beta level), measures the probability that we do not reject the null hypothesis H_0 when it it false. For example, our data indicates that the coin could still be fair, when in fact it is a biased coin. The probability that our test does not incur a type II error gives the *power of the test.*

Examples:

1. We test if a coin is fair by finding the proportion of heads. The null hypothesis is $H_0 =$ "$p = 50\%$."
   - A type I error is caused when the estimate for the proportion of heads is not close to 50%, so we reject H_0, but the coin is fair.
   - Making a type II error means that the estimate for the proportion of heads is close to 50%, but the coin is biased.
2. We devise a medical test to detect a disease. The null hypothesis is $H_0 =$you have the disease.
   - A type I error is caused when the test returns negative even if the patient has the disease.
   - Making a type II error means that our medical test returns positive, but we do not have the disease.
3. A person is put on trial and assumed to be innocent until proven guilty. The null hypothesis is $H_0 =$person is innocent.
   - Making a type I error means that the defendant is convicted, but he is innocent.
   - A type II error is caused when the defendant is set free, but he is guilty.
4. We test if a message is spam. The null hypothesis is $H_0 =$ message is spam.
   - A type I error is caused when a spam message is not placed in the junk folder.
   - Making a type II error means a good message was placed in the junk folder.

One thing to keep in mind is that one would like to minimize both types of error, but that is not possible. In fact, the smaller level α for type I error we allow, the larger will be the probability β of not rejecting a null hypothesis when false. Which error is more important to minimize depends on our experiment and what we are trying to measure. The following table summarizes the outcomes of hypothesis testing:

	H_0 is True	H_0 is False
Reject H_0	Type I error	OK
Not reject H_0	OK	Type II error

How the test works:

- Decide on the level of significance (how much error of making the wrong decision are we willing to tolerate).
- Assume H_0 is true and the parameter tested has the value given in the null hypothesis.
- Sample the parameter by computing a sample statistic s.
- Find the z-score for the observation of s, by subtracting the mean and dividing by the standard deviation (these are dependent on the parameter from H_0).
- Compute the ***p*-value** for the z-score, that is, find the probability that we observe data as extreme as our sample, by assuming H_0 is true.

Suppose we estimate q and we want to test

$$H_0 : "q = a" \text{ ,for some fixed value } a.$$

We compute the z-score for the sample statistic. Suppose the z-score we obtain equals b. We consider the following alternative hypotheses and the related p-values.

$$\begin{array}{lcllcl}
H_1 & : & "q \neq a" & \Rightarrow & p\text{-value} = P(z < -|b|) + P(z > |b|) & \text{two-tailed } p\text{-values} \\
H_2 & : & "q < a" & \Rightarrow & p\text{-value} = P(z < b) & \text{left-tailed } p\text{-values} \\
H_3 & : & "q > a" & \Rightarrow & p\text{-value} = P(z > b) & \text{right-tailed } p\text{-values}
\end{array}$$

Once we find the p-value, decision about the hypothesis follows in the same way, as in the next step.

- Decide to reject or not reject H_0 in favor of the alternative hypothesis based on

$$\boxed{\text{if } p\text{-value} < \text{level of signififcance} \Rightarrow \text{reject } H_0 \text{ in favor of alternate hypothesis}}$$

otherwise do not reject H_0 in favor of alternate, in which case we say we *failed to reject the null hypothesis at α level, or with confidence* $1 - \alpha$.

Example: We test if a coin is fair by tossing it 400 times. Let p be the probability that the coin lands on heads. Let the null hypothesis $H_0 : "p = 0.5"$ be tested against the alternate $H_1 : "p \neq 0.5"$.

(a) Note that $\dfrac{\#\text{ heads}}{400}$ has mean 0.5 and variance $\dfrac{(0.5)(1-0.5)}{400} = 0.000625$

(b) We decide on a level of significance $\alpha = 0.05$.

(c) Suppose we obtain 216 heads, so the proportion of heads is $s = 0.54$

$$z\text{-score} = \frac{0.54 - 0.5}{0.025} = 1.6$$

$$p\text{-value} = P(z < -1.6) + P(z > 1.6) = .1096 > 0.05.$$

Since the p-value is larger than the α-level, we do NOT reject H_0.

(d) Suppose we obtain 160 heads. Then the proportion of heads is $s = 0.4$

$$z\text{-score} = \frac{0.4 - 0.5}{0.025} = -4$$

$$p\text{-value} = P(z < -4) + P(z > 4) = .0002 < 0.05.$$

Since the p-value is smaller than the α-level, we reject H_0 in favor of H_1.