MAT 105 - Summary of Lecture 20
Thursday, March 24, 2016

> *When you have eliminated the impossible, what ever remains, however improbable, must be the truth.*
> (Sir Arthur Conan Doyle)

Estimation

Suppose we want to estimate a quantity: the proportion of people who vote Democrat in an election, the proportion of working components in a system, the percentage of people who like a certain feature in a game, the sum of a large number of dice rolls, the average value from a die roll, etc. How do we estimate the quantity and how do we measure how accurate is our estimate?

- first we need to get enough measurements (large amount of data). Suppose we take n measurements (data points) $x_1, x_2, \ldots, x_n$.
- the data will be distributed according to a bell curve (the more data, the closer to a bell curve)
- the expectation μ can be computed by taking a simple average (follows from the Law of Large Numbers)

$$\mu \approx \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

- the variance σ^2 can be derived from the random variable if we know its type (Binomial, Bernoulli, Geometric etc), otherwise we compute the variance as in the beginning of the semester. That is,

$$\sigma^2 = \frac{(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}.$$

- We normalize the data set and use the Central Limit Theorem to find probabilities related to this data.
- We analyze the data and estimate the error in our measurement of the desired quantity.
- If we wanted to check a hypothesis about the quantity (such as $p = 0.5$ for a coin to land on heads), we make a decision on the hypothesis.

Confidence intervals (cont.)

A way to assess the accuracy of our measurement, based on a sample, is by looking at information provided by the Central Central Limit Theorem. Recall that for data distributed according to a normal distribution, or that can be approximated by a normal distribution, we can find confidence intervals. Confidence intervals can be interpreted in two ways. For example, for the 90% confidence interval,

- 90% of the data falls in the 90% interval
- the probability that a data point is in the 90% interval is 0.9.

The most commonly used confidence intervals are the 95% confidence interval and 99% confidence interval.

- the 95% confidence interval for a $N(\mu, \sigma^2)$ is $[\mu - 2\sigma\,,\, \mu + 2\sigma]$
- the 99% confidence interval for a $N(\mu, \sigma^2)$ is $[\mu - 3\sigma\,,\, \mu + 3\sigma]$

Remarks: Note that for more confidence, the interval is larger. That is, the more error you allow, the more certain you can be that the true value of the quantity measured is included in the interval. Also, the error in the measurement, or **the margin of error**, is given by the half width of the confidence interval. More precisely, the margin of error for 95% confidence is 2σ, while the margin of error for 99% confidence is 3σ.

Example: We poll 900 people and estimate the proportion who vote Democrat. Suppose we find out that a proportion p voted Democrat.

(a) What is the 95% confidence interval for the outcome of the election?

(b) What is the 99% confidence interval for the outcome of the election?

(c) What is the error in measurement for the 95% confidence interval?

(d) How many people should we poll for a 1% error in a 95% confidence interval?

Answers: Set $X_i = 1$ if the ith person polled voted Democrat, and 0 otherwise. We are looking for confidence intervals for the proportion

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_{900}}{900}.$$

We know that $E[\overline{X}] = E[X_1] = p$ and $Var(\overline{X}) = \dfrac{Var(X_1)}{900} = \dfrac{p(1-p)}{900}$, so $StDev(\overline{X}) = \dfrac{\sqrt{p(1-p)}}{30}$.

(a) The 95% confidence interval is $\left[p - \dfrac{2\sqrt{p(1-p)}}{30}\,,\, p + \dfrac{2\sqrt{p(1-p)}}{30}\right]$.

(b) The 95% confidence interval is $\left[p - \dfrac{3\sqrt{p(1-p)}}{30}\,,\, p + \dfrac{3\sqrt{p(1-p)}}{30}\right]$.

(c) The error in our measurement, with 95% confidence, is at most

$$\frac{2\sqrt{p(1-p)}}{30} \leq \frac{2 \cdot (1/2)}{30} = .033$$

since the maximum $\sqrt{p(1-p)}$ can be is $1/2$. In polling results, this is stated as "a margin of error of $\pm 3\%$".

(d) We want to poll n people and our result, in the worst case scenario when $\sqrt{p(1-p)} = 1/2$, should have an error no larger than 1%. That is,

$$\frac{2\sqrt{p(1-p)}}{\sqrt{n}} \leq 0.01 \;\Leftrightarrow\; \frac{1}{\sqrt{n}} \leq 0.01 \;\Leftrightarrow\; \sqrt{n} \geq 100 \;\Leftrightarrow\; n \geq 10,000.$$

Thus, we should poll at least 10,000 people for an error this small.

MAT 105 - Group Work
March 24, 2016

I. Suppose $X_1, X_2, \ldots$ are random variables with mean $\mu = 3$ and variance $\sigma^2 = 4$. Let $S_{100} = X_1 + \cdots + X_{100}$ be the sum of the first 100 such random variables.

(a) Approximate $\dfrac{S_{100}}{100}$

(b) Find the probability $P(260 \leq S_{100} \leq 340)$

II. We want to compute the **proportion** of heads in 100 fair coin flips.

(a) Find the probability that the proportion is less than 60%?

(b) Find the probability that the proportion is less than 40%?

III. You want to check if a six-sided die is fair. You decide to count the number of times 6 comes up. Let S_n count the number of 6's in n rolls and let p be the probability that 6 comes up in a roll. If the die is indeed fair and $p = 1/6$,

(a) approximate $\dfrac{S_n}{n}$ for n large.

(b) find the 95% confidence interval for S_{100}

(c) find the 95% confidence interval for $\dfrac{S_{100}}{100}$

(d) find the 99% confidence interval for $\dfrac{S_{100}}{100}$

(e) what is the margin of error for your measurement in (d)?