

Quantifying Chance

Part 2: Understanding Chance

INFO-1301, Quantitative Reasoning 1
University of Colorado Boulder

April 5, 2017

Prof. Michael Paul

Sampling Distribution

The sampling distribution is approximately normal

- The mean is the true population mean
- The standard deviation is called the **standard error (SE)**

$$SE = \frac{\sigma}{\sqrt{n}}$$

This is known as the
Central Limit Theorem

- σ is the standard deviation of your data (unknown – so use the standard deviation from your sample)
- n is the size of your sample
 - Larger $n \rightarrow$ smaller standard error
(sample mean is more likely to be close to population mean)

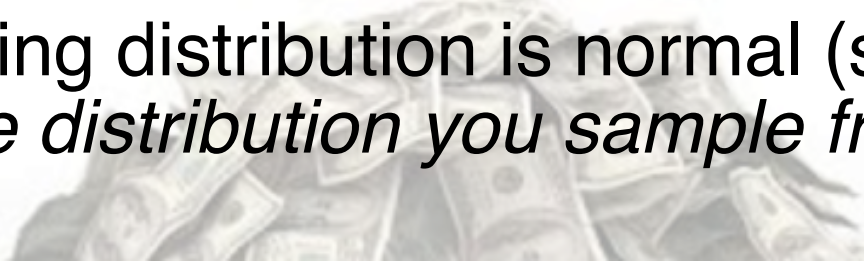
A Visualization

- <http://students.brown.edu/seeing-theory/statistical-inference/index.html#first>

An Example

- US household income is heavily right-skewed
 - (thanks to people like Bill Gates and Warren Buffett)
 - Can tell this from the large difference between median = 51.9 (K\$) and the mean = 71.9 (K\$)
- Even though the population is right skewed, when you take 1000 samples of size $n=100$, they will form a normal distribution around 71.9.

A property of the Central Limit Theorem:
the sampling distribution is normal (symmetric),
even if the distribution you sample from is not!



An Example

- US household income is heavily right-skewed
 - (thanks to people like Bill Gates and Warren Buffett)
 - Can tell this from the large difference between median = 51.9 (K\$) and the mean = 71.9 (K\$)
- Even though the population is right skewed, when you take 1000 samples of size $n=100$, they will form a normal distribution around 71.9.
- If you took 1000 samples of size $n=25$, the sample means would form a normal distribution with twice the standard error as $n=100$.



An Example

- If you changed your population to single-earner households where the employed person was a public school teacher, the population would be less dispersed and thus the sample means would also be less dispersed ($\mu=56.3$ in 2014, NCES)
- If you surveyed a sample of 100 people and the sample mean of their salary was 82.1, you would know either that this sample was not drawn from public school teachers or that it was not a random sample of teachers because 82.1 is far removed from 56.3.
 - How far removed?
 - If $SD = 18.0$, then $SE = 18.0 / 10 = 1.8$
 $Z = (82.1 - 56.3) / 1.8 = 14.33$
 - The probability of a sample mean this extreme is < 0.00001

P-Values

In the previous example, we said that if our sample mean was 82.1, we must not have sampled from public school teachers.

The reason we can be confident about this is that it is extremely unlikely we would have gotten this mean if we randomly sampled from teachers.

- There is a very small probability (< 0.00001) that we could have gotten this measurement.

In the context of describing your confidence of a measurement, this probability is called a **p-value**.

P-Values

Loosely speaking, a **p-value** is the probability of getting a particular measurement by chance.

- If your p-value is very small, then you most likely need to come up with a different explanation for your result than your default assumption
 - If you are testing a new theory, a low p-value for the old idea is evidence that the new idea is true

A more rigorous explanation of p-values can be found in Diez 4.3 (but not required)

Another Example

Gloria Ann Howland

- Gloria Ann Howland

Another Example

Statistician Charles Sanders Pierce examined 42 genuine signatures, and looked to see how often the strokes in pairs of signatures lined up

- He examined all 861 pairs of signatures
 - Where does this come from? $42 \text{ choose } 2 = 861$

The strokes were a match one-fifth of the time

- Probability of a match is 0.2

The signature in question contained 30 strokes

- What is the probability of all 30 strokes matching?
 $0.2^{30} = 0.00000000000000000000000001$
 - This assumes strokes are independent, which isn't quite right.

Another Example

If we assume that this is a good probability model for signatures, then the p-value of having an exact match in this case is 0.000000000000000000000001.

In other words, it is extremely unlikely that the signatures would be an exact match by chance.

- There is likely some other explanation (e.g., forgery)

Summary

P-values can tell you if an observation, or the differences between observations, are meaningful

- Can detect fraud or cheating
- Can say if an experiment has an effect that is significant (e.g., does bacon cause cancer?)

Summary

If a p-value is **large**, that means it is likely what you observed is due to chance, so it is unlikely to be meaningful.

If a p-value is **small**, that means it is unlikely to be due to chance, so there is likely another explanation.

- What is considered small? It depends, but a commonly accepted cutoff is **$p < 0.05$**