STA130 W4

Wednesday, September 11, 2019 3:02 PM

Statistical Inference
Population - what we interested in but scope may be to large(cannot observe it directly in general)
Sample - smaller, we want to use info from the sample to answer question about the population
Sample **Inference** the population, the process is called **sampling**
Sampling - Random(ideally), so that sample is representative of the population.

- An **inference** is an uncertain conclusion or generalization
- **Statistical inference** helps us make conclusions(generalization) or decisions based on statistical information(subject to randomness and uncertainty)
  i.e. conclude population using random sample
- Conclusion is uncertain but we try to measure the uncertainty

Sampling variability of estimated proportion of heads(p hat)
P: the probability of getting heads
P-hat: the (estimated) proportion of heads in 10 coin flips(sample)
- Symmetric distribution; 0.5 has the highest p-hat

$\hat{p}$

Sampling distribution
P-hat means it is calculated from a sample
P: parameter value in the population
- Need many samples of 10 coin flips

The **sampling distribution of a statistic**(i.e. p-hat histogram) is the distribution of statistic values taken for all possible samples of the same size(n, i.e. n=10) from the same population

A **simulation** is a way to explore random events(what the data could look like under certain assumptions)
R code:

```
Sample(c("heads", "tails"),
        Size=10,
        Prob=c(0.5, 0.5),
        Replace=TRUE)
#prob = #outcome
```

First argument: a vector from which we want to sample
Replace: TRUE = sampling **with** replacement, FALSE = sampling without replacement(default)

```
Coin <- c("heads", "tails")
Flips <- Sample(coin,
                Size=10,
                Prob=c(0.5, 0.5),
                Replace=TRUE)
```

*Flips
*OR: table(flips)

```
Sum(flips=="heads")/10  == mean(flips=="heads")

Sim <- data_frame(p_heads = mean(flips=="heads"))
Sim %>% ggplot(aes(x=p_heads)) +
        Geom_dotplot() + xlim(0, 1) + ylim(0,10) +
        Labs(x="proportion heads in 10 coin flips")
```

*add simulation to build this distribution graph

For loops
Automate the process of generating simulations

```
For (i in 1:100)
{
        *code
}
```

*i= variable telling R which repetition we are on(any name)
*1000=the number of repetitions to go over

Reproduce randomness
Force the sample function to produce the same outcome every time by setting a parameter called the "seed".
R code:

```
Set.seed(130)
Sample(c("H","T"), size=10, prob= c(0.5, 0.5), replace = TRUE)
```

*130: has to be integer

Steps to estimate a sampling distribution:

1.Set values for simulation(size, n of repetitions, seed)

```
N_obs <- 10
Repetitions <- 1000
Simulated_stats <- rep(NA, repetitions)*creates a vector repeat NA 1000 times
Set.seed(101)*result don't change
```

2.Use for loop

```
For (i in 1:repetitions){
        New_sim <- sample(c("heads", "tails"),
                        Size = n_obs,
                        Prob = c(0.5, 0.5),
                        Replace = TRUE)
        Sim_p <- sum(new_sim == "heads")/ n_obs

        Simulated_stats[i] <- sim_p
}
```

*same sim_p in the ith slot of simulated_stats vector

3.Turn the result into a data frame

```
Sim <- data_frame(p_heads = simulated_stats)
```

4.Plot the results

```
Sim %>% ggplot(aes(x= p_heads)) +
        Geom_histogram(binwidth =0.1, colour ="black", fill = "gray") +
        xlab("Proportion of heads in 10 coin flips")
```

- Although the samples are random, the distribution has a specific distribution

Observe unusual/extreme results

P-hat **unusual**/**extreme** compared to this distribution or is it **consistent** with the distribution?

1. Question : do they behave like regular coin (p=0.5)
2. P-hat = 0(from obs)
3. Estimated sampling distribution using regular coin(p=0.5) using simulation
4. Compare& conclude

->very usual, seems our initial assumption was false

*repetition does not affect the shape but the **precision**

Significance testing(hypothesis testing): one type of statistical inference

Sometimes statistical inference is not appropriated(if we have all individuals in the population)

Steps for conducting a hypothesis test for a proportion p

1. State **hypothesis** (Ho and Ha)

   Null hypothesis (**Ho***nothing going on) Couples are equally likely to tilt to right or left.

   Ho: p=0.5 default value

   Alternative hypothesis(Ha/HA/H1): p≠0.5

   P is the proportion of individuals tilt to right when they kiss

2. Calculate the **test statistic** based on observed data

   A **parameter** is a number that describes the population. It is the "true" value of what we're interested in, for the population we are focused on.

   A **statistic** is a number that describes the sample. The value of a statistic will changes from sample to sample (ex: sample mean, median, variance, etc)

   A **test statistic** is a special statistic that helps us decide whether the data is compatible with **Ho**.

   One-sided test: if you are looking at one is higher than the other(Ho>=5)

   In this example:

   Parameter: **p=?**: the true proportion of people who kiss to the right (population = all couples)

   Statistic: Ex **p-hat=80/124=0.645** is the proportion of people who kiss to the right. This value would likely be different if we got a different sample of 124 couples.

   The **test statistic** is a number, calculated from the data. For the kissing example, the test statistic we'll use is **p-hat=80/124=0.645** since it is the sample version of the parameter we're interested in

3. Simulate samples under Ho and calculate the statistic for each sample

   Assuming Ho is true and calculate the statistic for each sample

   Model what we would expect to see if couples had no preference

   Simulate, and For each sample calculate the proportion tilt their head to the right, repeat

```
n_observations <- 124
repetitions <- 1000
simulated_stats <- rep(NA, repetitions)
set.seed(101)
for (i in 1:repetitions){
  new_sim <- sample(c("right", "left"),
                  size = n_observations,
                  prob = c(0.5,0.5),
                 replace = TRUE)
  sim_p <- sum(new_sim == "right") / n_observations

  simulated_stats[i] <- sim_p;
}
```

```
sim <- data_frame(p_right = simulated_stats)
sim %>%
 ggplot(aes(x = p_right)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey") +
  xlab("Simulated proportions of individuals who kiss to the right if p=0.5 (samples of size n=124)")
```

***More simulated values implies better estimate** of the sampling distribution for our statistic
In practice, the number of simulations is more typically on the order of 10,000.
Using a computer for simulation: at least 1,000, ideally 10,000

In the observed sample (not simulated!) we observed that **p-hat=80/124=0.645**, so 64.5% of couples kissed to the right.

4. Evaluate the evidence against Ho

The **P-value** is the probability of observing data that are **at least as unusual** (or **at least as extreme**) as the sample data, under the assumption that it is true.
We estimate the P-value as the proportion of values in the estimated sampling distribution that are as extreme or more extreme than the test statistic calculated from our observed sample data.
For the kissing example:
Null hypothesis value: p=0.5
Observed estimate from the sample: p = 0.645
Values at least as extreme/unusual as the sample statistic: all values **greater or equal to 0.645** and **all values less than or equal to 0.5 - (0.645 - 0.5) = 0.355**
i.e. values further away from the null value (p=0.5) than the test statistic is
(i.e. further than |0.645-0.5|away from p=0.5)
This is a **two-sided test** because it considers differences from the null hypothesis that are both larger and smaller than what you observed.

Find the proportion of simulated values that are at least as unusual as p-hat=0.645

```
sim %>% ggplot(aes(p_right)) +
 geom_histogram(binwidth = 0.02,
 colour = "black", fill = "grey") +
 geom_vline(xintercept = 0.645, color = "red") +
  geom_vline(xintercept = 0.355, color = "blue") +
 labs(x = "Simulated proportions of individuals who kiss to the right if p=0.5 (samples of size n=124)")
```

Find P-value

```
pvalue <- sim %>%
   filter(p_right >= 0.645 | p_right <= 0.355) %>%
   summarise(p_value = n() / repetitions)

as.numeric(pvalue)
```

OR

```
pvalue <- sim %>%
  filter(abs(p_right - 0.5) >= abs(0.645 - 0.5)) %>%
  summarise(p_value = n() / repetitions)

Pvalue
```

A **small p-value** tells us that there is only a small chance that we would observe a test statistic as far away from the null value of the parameter if were really true
Two reasons that can lead to a small p-value:
1.Ho is actually true and we just observed an unlikely extreme value of the statistic
2.Ho is not true
The smaller the p-value, the more we lean towards (2) - in other words, the smaller the p-value, the more "evidence" we have against Ho

5. Make a **conclusion**

strength of evidence against Ho

P-value	Evidence
p-value > 0.10	no evidence against
0.05 < p-value < 0.10	weak evidence against
0.01 < p-value < 0.05	moderate evidence against
0.001 < p-value < 0.01	strong evidence against
p-value < 0.001	very strong evidence against

Statistical significance - the likelihood of the outcome(5% in this case)

Conclusion

A significance level (α) set in advance determines the cut-off for how unusual/extreme the test statistic has to be (assuming is true) in order to reject the assumption that is true (i.e. to conclude statistical significance)

α can be chosen to be any number but typically α =0.05
RULE: Reject Ho if p-value ≥ α

It is better to report the p-value and comment on the strength of evidence against Ho instead of only reporting whether the result is/isn't statistically significant

Since the P-value is 0.001 we conclude that we have we have strong evidence against the null hypothesis that

individuals have no preference for kissing to the left or to the right.
The data provide convincing evidence that people are **more likely** to tilt their heads to one direction when they kiss, and suggests a preference for tilting to the right.