Chapter 13 Notes

Sampling Distributions
STP-231

Arizona State University

Parameters vs Statistics

- Statistics desribe sample characteristics, $\overline{Y}$ and S are sample mean and sample standard deviation
- Parameters describe population characteristics. Usually trying to estimate these
- Proportions: are a relative frequency, can be for population or sample

Measure	Statistic	Parameter
Proportion	$\hat{p}$	p
Mean	$\overline{Y}$	μ
Standard deviation	S	σ
Variance	S^2	σ^2

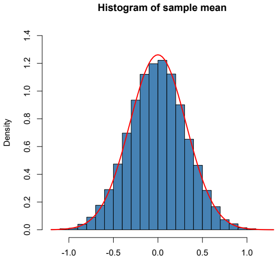
Sampling

- We conduct an experiment on a random sample, hoping to be representative of a population
- Cannot expect sample to perfectly approximate population
- Statistics is distinguishing whether differences between sample and population are from random chance or if there is a real effect
- **Sampling error**: discrepancy between the sample and population

Sampling (continued)

If a sample is chosen randomly:

- Statistics themselves have their own distribution, i.e. $\hat{p}$ and $\overline{Y}$ are random variables themselves
- For example, the picture below is the distribution of the sample mean of a random sample from a normal distribution ($\overline{Y} \sim N(\mu_Y, \sigma_Y^2/n)$)



Sampling (continued)

If a sample is chosen randomly:

- We can interpret the sampling distribution of a statistic as “what would happen if we sample many times”, though this is often not feasible
- Simulation is one solution, we can use software to imitate the random behavior

Sampling Distribution of a statistic

- Distribution of all of the possible observations of a statistic for samples of a given size from a population

- Ideal pattern that would emerge if we could view all samples of a given size. Shape, center, and spread are still useful descriptive properties

- The goal is to see how closely does our sample resemble the population

Notation

- Population parameters will either be subscripted with the random variable in question, “pop”, or not at all, i.e.

  $$\mu_Y \text{ or } \mu_{\text{pop}} \text{ or } \mu$$

  same deal with σ

- The parameters of the sampling distribution of any statistic will be subscripted with the statistic in question, i.e. $\mu_{\overline{Y}}$ or $\sigma_{\overline{Y}}$ if we are looking at the sample mean distribution, or $\mu_{\hat{p}}$ if looking at sample proportion for example

- The sampling distribution of the mean is $N(\mu, \sigma/\sqrt{n})$ when the underlying population is normal

Meta Study

- Experiments where every possible sample that can be taken from a population is indeed taken
- Collecting data from every sample is unlikely
- Theoretical experiments, not often done in practice

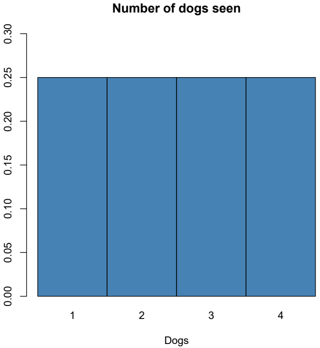
Example 1: Meta Study

Imagine we are part of a class project in a veterinary class. We want to find out how many pups ASU students have. Unfortunately, we cannot ask every student, so we merely ask 2. Assume

- The actual number of dogs follows a uniform(1,4) distribution (same probability for each whole number 1,2,3, or 4)
- Let X be the # of dogs people have, where $\Pr(X = 1) = \Pr(X = 2) = \Pr(X = 3) = \Pr(X = 4)$
- This means nobody has 0 dogs, and nobody has more than 4 (assume this is population truth)

Population Characteristics

$$\mu_X = 2.5 \qquad \sigma_X = \sqrt{\frac{15}{12}}$$



Possible Samples of Size $n = 2$

- If we choose 2 people and ask how many dogs they saw, we have 4^2 possible combinations answers they give us. This is because here we sample with replacement (i.e. an observation can be chosen again once its in the sample). This is so we can find a probability of that specific combination

- The 16 possible samples:

$$\{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(2,4),(3,1),(3,2),(3,3),(3,4),(4,1),(4,2),(4,3),(4,4)\}$$

Sampling Distribution

The order of the data for each sample is negligible, i.e. (1,2) is the same as (2,1)

Sample	Pr(Sample)
1 and 1	0.0625
2 and 1	0.125
3 and 1	0.125
4 and 1	0.125
2 and 2	0.0625
3 and 2	0.125
4 and 2	0.125
3 and 3	0.0625
4 and 3	0.125
4 and 4	0.0625

Sampling Distribution of $\overline{X}$

Sample	Sample Mean
1 and 1	1
2 and 1	3/2
1 and 3; 2 and 2	2
1 and 4; 2 and 3	5/2
2 and 4; 3 and 3	3
4 and 3	7/2
4 and 4	4

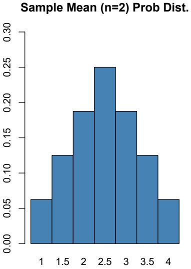
Sampling Distribution of $\overline{X}$ Continued

Sample Mean	Pr(Sample Mean)
1	0.0625
3/2	0.125
2	0.1875
5/2	0.25
3	0.1875
7/2	0.125
4	0.0625

Example:

$$\begin{aligned}\Pr\left(\overline{X}=2\right) &= \Pr(\text{sample with observations “2” and “2”}) \\ &+ \Pr(\text{sample with observations “1” and “3”}) = 0.1875\end{aligned}$$

Summary Measures of Sampling Distribution of $\overline{X}$



$$\mu_{\overline{X}} = 2.50 \qquad \sigma = \sqrt{0.67} = 0.81 \approx \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

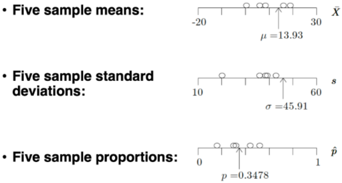
where $n = 2$. However, using a normal approximation implies we'd have a sample mean of 4.1 dogs or higher in 2.5% of our samples. Not the best!

Example 1: Conclusions

- We constructed the sampling distribution of the mean by exhausting all possible samples within our study.
- The sampling distribution mean was equal to the mean, and the standard deviation was a function of the population standard deviation
- The sample distribution appears approximately normal. If you were to choose multiple random samples from the population, the sample mean would be exactly the population mean 25% of the time.

Example: Variability of Random Samples

Dot plots of sample statistics from 5 random samples of size $n = 25$.



- Five sample means:
- Five sample standard deviations:
- Five sample proportions:

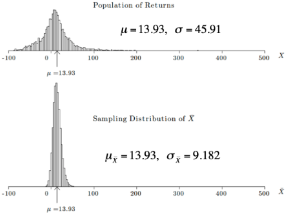
Notice the placement of the pop mean, pop proportion, and pop standard deviation relative to stats

Takeaways

- Statistics vary from sample to sample
- Statistics of samples are themselves random variables and have distributions. This allows us to generalize our results without taking every possible sample, of which there are a lot
- We can answer questions like how close to the true μ is $\overline{X}$ likely to be?
- What is the shape, center, and spread of the distribution?

Example 2

Take 1,000,000 samples of size $n = 25$ and plot the distribution of the resulting sample averages. Notice, the mean location is the same, but the variance is tighter. Notice, the number of samples is there to show the behavior is true, but the variance only depends on sample size!



Properties of sampling distribution of $\overline{X}$

- Assume our population is normally distributed. Then:
- The mean of the sampling distribution of $\overline{X}$ is the populetion mean, i.e. it's unbiased
- The spread of the sampling distribution of $\overline{X}$ is given by the standard deviation, which is

$$\sigma_{\overline{X}} = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

Central Limit Theorem

- In fact, for any distribution as the sample gets larger:
- The standard deviation of the sampling distribution decreases
- The sampling distribution becomes normal, specifically of the sample mean
- n=10 pretty well usually CLT viz

Central Limit Theorem Continued

- Regardless of shape, $n = 10$ provides an approximately normal sampling distribution
- A distribution tends to look very normal for $n = 30$
- Regardless of normaility of the population, the sampling distribution of the sample mean will be normal provided n is large enough

Example 3

Phone call lengths are distributed normally with $\mu = 8$ and $\sigma = 2$ minutes. If you select a random sample of 25 calls, what percentage of the sample means would be between 7.8 and 8.2 minutes?

Example 3

Phone call lengths are distributed normally with $\mu = 8$ and $\sigma = 2$ minutes. If you select a random sample of 25 calls, what percentage of the sample means would be between 7.8 and 8.2 minutes?

With a sample of 25, we'd expect a sampling distribution of the mean to be

$$N(8, 2/\sqrt{25}) = N(8, 0.40)$$

So about 38% of the data would be in this range.

Sampling Distribution of $\hat{p}$

- $\hat{p}$ is the proportion of observations that satisfy a condition within a sample

- Let X be the count of the occurrences of some categorical outcome in a fixed number of observations. Let n be the number of observations. Then the sample proportion is

$$\hat{p} = \frac{X}{n}$$

Sampling Distribution of $\hat{p}$

- Choose an SRS of size n from a large population that contains population proportion p of successes. Then:

- $\hat{p} = \frac{\text{\# of successes in a sample}}{\text{sample size}}$

- The mean of the sampling distribution if p, so its unbiased. nice

- The standard deviation of the sampling distribution if $\sqrt{p(1-p)/n}$. As the sample size increases, the sampling distribution $\hat{p}$ becomes approximately normal

Sampling Distribution of Variance

- The sample variance S^2 has a mean of σ^2.
- We'll see this later, but $(n-1)S^2/\sigma^2$ is distributed χ^2_{n-1}, which will be useful later on in the semester

Law of Large Numbers

- The standard deviations of the sampling distributions are functions of sample size

$$\sigma_{\hat{p}} = \sqrt{p(1-p) \cdot n}$$
$$\sigma_{\overline{X}} = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

- If we continue increasing n then the variability in the distributions will be reduced

- Reduce in spread means statistics will get closer and closer to their respective parameters,

$$\overline{x} \rightarrow \mu_X$$
$$\hat{p} \rightarrow p$$

Our pupper friend