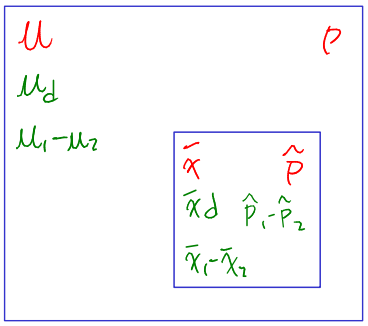
Probabilities for the Big 5

David Armstrong

UCI

To Learn About A Parameter



- Collect data appropriate for that parameter
- Calculate a sample statistic

  Sample mean

  Sample proportion
- Perform Statistical Inference

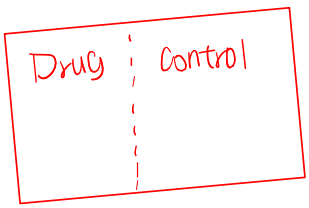
  Confidence Interval-Gives an interval of likely values for your parameter

  Hypothesis Test-Use your data to make a decision about your parameter

The Big Five

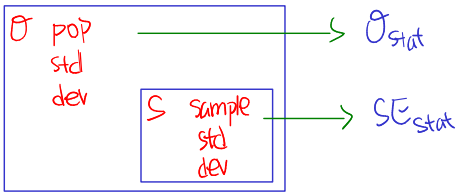
Type Of Variable	Parameter Description	Population Parameter	Sample Statistic
Categorical	1 population proportion	p	$\hat{p}$
Categorical	Difference in 2 population proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
Quantitative	1 population mean	μ	$\bar{X}$
Quantitative	Population mean of paired differences (dependent samples)	μ_d	$\bar{X}_d$
Quantitative	Difference in 2 population means for independent samples	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$

matched Pairs



General Format for Sampling Distributions

Sample Statistic $\sim$ Approximately Normal($\mu_{Statistic} = E[Statistic], \sigma_{Statistic} = SD[Statistic]$)



$$Z = \frac{Statistic - \mu_{Statistic}}{\sigma_{Statistic}}$$

If we do NOT know the $SD[statistic]$ we can estimate it with the standard error $SE[statistic]$.

$\hat{\sigma}_{statistic} = SE[statistic]$

- The sample statistic ESTIMATES the population parameter

  Take m random samples and calculate the sample statistic for each of the m samples. Then the average of the m sample statistics is the expected value of the sample statistic $= \mu_{statistic}$. If m is large, then $\mu_{statistic}$ will be equal to the population parameter.

- The standard error of the sample statistic ESTIMATES the standard deviation of the sample statistic

  They both measure the typical spread of the sample statistic about the mean.

$SS \sim ANL(\mu_{SS}, \theta_{SS})$ if $\theta_{SS} = ?$ then use SE

The Big Five

Parameter Description	Population Parameter	Sample Statistic	Expected Value	Standard Deviation	Standard Error
1 population proportion	p	$\hat{p}$	p	$\sqrt{\frac{pq}{n}}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$
Difference in 2 population proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$	$\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$
1 population mean	μ	$\bar{X}$	μ	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
Population mean of paired differences (dependent samples)	μ_d	$\bar{X}_d$	μ_d	$\frac{\sigma_d}{\sqrt{n_d}}$	$\frac{s_d}{\sqrt{n_d}}$
Difference in 2 population means for independent samples	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Sampling Distributions: Proportions

$$\hat{p} \sim \text{Approximately Normal}\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}\right)$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$\hat{p} \sim \text{Approximately Normal}\left(\mu_{\hat{p}} = p, SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

$$\hat{p}_1 - \hat{p}_2 \sim \text{AppNormal}\left(\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2, \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$$\hat{p}_1 - \hat{p}_2 \sim \text{AppNormal}\left(\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2, SE[\hat{p}_1 - \hat{p}_2] = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}\right)$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

Important Vocabulary

Write out the notation for the following list:

population proportion p
sample proportion $\hat{p}$
standard deviation of the sample proportion $\sigma_{\hat{p}}$
standard error of the sample proportion $SE_{\hat{p}}$
population proportion for group 1
sample proportion for group 1
population proportion for group 2
sample proportion for group 2
sample size
sample size for group 1
sample size for group 2
difference between two population proportions
difference between two sample proportions
standard deviation of the difference between 2 sample proportions
standard error of the difference between 2 sample proportions

Sampling Distributions: Means

$$\bar{X} \sim \text{AppNormal}\left(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\right)$$

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\bar{X} \sim \text{t}\left(\mu_{\bar{X}} = \mu, SE[\bar{X}] = \frac{s}{\sqrt{n}}, df = n - 1\right)$$

Degree of freedom

$n - 1$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$\bar{X}_d \sim \text{AppNormal}\left(\mu_{\bar{X}_d} = \mu_d, \sigma_{\bar{X}_d} = \frac{\sigma_d}{\sqrt{n_d}}\right)$$

# of pairs

$n = 100$

$n_d = 100$

$$Z = \frac{\bar{X}_d - \mu_d}{\frac{\sigma_d}{\sqrt{n_d}}}$$

$n = 100$

$n_d = 100$

$$\bar{X}_d \sim \text{t}\left(\mu_{\bar{X}_d} = \mu_d, SE[\bar{X}_d] = \frac{s_d}{\sqrt{n_d}}, df = n_d - 1\right)$$

$$t = \frac{\bar{X}_d - \mu_d}{\frac{s_d}{\sqrt{n_d}}}$$

$$\bar{X}_1 - \bar{X}_2 \sim \text{AppNormal}\left(\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2, \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

E(x)

Obs val

E(x)

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\hat{\theta}_p$

$$\bar{X}_1 - \bar{X}_2 \sim$$
$$\text{t}\left(\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2, SE[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, df = min(n_1 - 1, n_2 - 1)\right)$$

Estimated

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Unpooled std dev $\theta_1 \neq \theta_2$

$$\bar{X}_1 - \bar{X}_2 \sim$$
$$\text{t}\left(\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2, SE[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, df = n_1 + n_2 - 2\right)$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

pooled std dev $\theta_1 = \theta_2$

Important Vocabulary

Write out the notation for the following list:

population mean μ
sample mean $\bar{x}$
population standard deviation σ
sample standard deviation s
standard deviation of the sample mean $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \rightarrow z$
standard error of the sample mean $SE_{\bar{x}} = \frac{s}{\sqrt{n}} \rightarrow t$
population mean of paired differences
sample mean of paired differences
standard deviation of the population of paired differences
standard deviation of the sample of paired differences
standard deviation of the sample mean of paired differences
standard error of the sample mean of paired differences
population mean for group 1
sample mean for group 1
population mean for group 2
sample mean for group 2
difference between 2 population means
difference between 2 sample means
population standard deviation for group 1
sample standard deviation for group 1
population standard deviation for group 2
sample standard deviation for group 2
standard deviation of the difference between 2 sample means
standard error of the difference between 2 sample means

EX: Suppose that out of all Olympic athletes, 30% of them train for more than 40 hours per week. Suppose a researcher took a sample of 250 athletes.

$p = 0.30$ $\quad n = 250$

(a) What proportion of the sample would be expected to train for more than 40 hours per week?

$E(\hat{P}) = p = 0.30$

(b) What is the sampling distribution of the sample proportion?

$\hat{P} \sim AN(\mu_{\hat{P}} = p = 0.30, \sigma_{\hat{P}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.30(0.70)}{250}} = 0.0290)$

(c) What is the probability that more than 35% of our sampled athletes train for more than 40 hours per week?

$P(\hat{P} > 0.35) = 1 - P(\hat{P} < 0.35)$

$= 1 - pnorm(0.35, 0.30, \sqrt{\frac{0.30(0.70)}{250}})$

$= 0.0422$

MEN
$n_1 = 200$
$\hat{p}_1 = 0.23$

WOMEN
$n_2 = 150$
$\hat{p}_2 = 0.16$

EX: Have you cheated? On the basis of surveys done in the past, men are more likely to cheat on exams compared to women. In a random sample of 200 men and 150 women, the sample proportion of men who cheated on an exam was 0.23 and the sample proportion of women who cheated on an was 0.16. Suppose it is known that about 10% more men cheat compared to women.

$p_1 - p_2 = 0.1$

(a) What do we expect the difference between the proportions to be for men compared to women?

$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0.10$

(b) Based on the information can we find the standard deviation of the difference between the sample proportions ?

$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1 q_1}{n} + \frac{pq}{n}} = \text{No}$

(c) What is the standard error for the difference in the two sample proportions?

$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} = \sqrt{\frac{0.23(0.77)}{200} + \frac{0.16(0.84)}{150}} = 0.0422$

(d) What is the probability that we see a difference smaller than the one we observed in our sample of men and women?

$\hat{p}_1 - \hat{p}_2 \sim AN(\mu_{\hat{p}_1 - \hat{p}_2} = 0.10, SE_{\hat{p}_1 - \hat{p}_2} \approx 0.0422)$

$P(\hat{p}_1 - \hat{p}_2 < 0.07) = \text{pnorm}(0.07, 0.10, \sqrt{\frac{0.23(0.77)}{200} + \frac{0.16(0.84)}{150}})$

$= 0.2386$

$n = 4 \quad \mu = 4 \quad \sigma = 1.2$

EX: A store manager is trying to decide whether to price bananas by weight, with a fixed cost per pound, or by the piece, with a fixed cost per banana. He is concerned that customers will choose the largest ones if there is a fixed price per banana. For one week the bananas are priced by the piece rather than by weight, and during this time the mean weight of the bananas purchased is recorded for all customers who buy 4 of them. The manager knows the population of weights of individual bananas is bell-shaped with mean of 4 ounces and a standard deviation of 1.2 ounces.

(a) What is the distribution of average weight of 4 bananas that each customer chooses?

$\bar{x} \sim AN(\mu_{\bar{x}} = \mu = 4, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{4}} = 0.60)$

(b) What is the probability that the average weight of the 4 bananas is greater than 6 ounces?

$$P(\bar{x} > 6) = 1 - P(\bar{x} < 6)$$
$$= 1 - \text{pnorm}(6, 4, 0.60)$$
$$= 0.0004$$

EX: Suppose at another store the population of weights of individual bananas is bell-shaped with mean of 4 ounces. Suppose we only have the value of the sample standard deviation; which is equal to 2.3.

$\mu = 4 \quad s = 2.3 \longrightarrow t$

(a) What is the distribution of average weight of 4 bananas that each customer chooses?

$\bar{x} \sim t(\mu_{\bar{x}} = \mu = 4, SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{2.3}{\sqrt{4}} = 1.15, df = n - 1 = 4 - 1 = 3)$

(b) What is the probability that the average weight of the 4 bananas is less than 3 ounces?

$$P(\bar{x} < 3) = P(t < -0.8696)$$
$$= pt(t, df) = pt(-0.8696, 3) = 0.2243$$

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{SE_{\bar{x}}} = \frac{3 - 4}{1.15} = -0.8696$$

Post - Pre

After - Before

EX: Suppose we have 10 football players that are selected to participate in a study. Although speed and strength are a necessity, flexibility and grace can also help their game. These players flexibility was measured in a sit and reach before taking Ballet classes for a month and then measured again at the end of the class. The Institute of Ballet claims that the average difference should be 4 inches. Below is a table of their scores: $\mu_d = \mu$

Before	12	6	7	8	9	10	11	15	3	5
After	13	12	10	9	10	8	10	15	9	8
	1	6	3	1	1	-2	-1	0	6	3

$\bar{x}_d = 1.8$ $\quad n_d = 10$

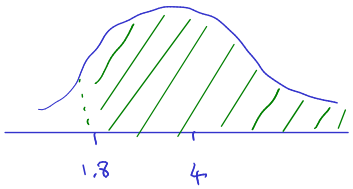
(a) What is the distribution of the sample mean of paired differences, if we know that the standard deviation is 5?

$\sigma_d = 5$

$$\bar{x} \sim ANC\left(\mu_{\bar{x}_d} = \mu_d = 4,\ \sigma_{\bar{x}_d} = \frac{\sigma_d}{\sqrt{n_d}} = \frac{5}{\sqrt{10}} \approx 0.5811\right)$$

(b) What is the probability that the average difference was greater than what we saw in our sample?

$$P(\bar{x} > 1.8) = 1 - P(\bar{x} < 1.8)$$
$$= 1 - pnorm\left(1.8, 4, \frac{5}{\sqrt{10}}\right)$$
$$= 0.9179$$



(c) Suppose we did not know the standard deviation. What is the probability that the average difference was greater than what we saw in our sample, if we know our sample standard deviation is 2.7?

$\sigma_d = ?$ $\quad S = 2.7$

$$\bar{x}_d \sim t\left(\mu_{\bar{x}_d} = \mu_d = 4,\ SE_{\bar{x}_d} = \frac{SD}{\sqrt{n_d}} = \frac{2.7}{\sqrt{10}} \approx 0.8538,\ df = n_d - 1 = 9\right)$$

$$t = \frac{1.8 - 4}{\left(\frac{2.7}{\sqrt{10}}\right)} = -2.5767$$

$$P(\bar{x} > 1.8) = P(t > -2.5767)$$
$$= 1 - pt(-2.5767, 9) = 0.9851$$

Treatment Control

EX: Suppose we have two groups of people that we would like to compare. The first group received a new weight loss drug. The second group thought they were receiving the drug, but instead were given a "sugar pill". The participants were weighed at the beginning of the study. After 4 weeks, the participants were weighed again. There weight loss was measured by subtracting their weight at the end of the study from their weight at the beginning of the study.

	Group 1 TR		Group 2 CO		
$\bar{x}_1$ ⟶	Observed average weight loss	10	Observed average weight loss	4	⟵ $\bar{x}_2$
s_1 ⟶	Standard Deviation	4	Standard Deviation	3	⟵ s_2
	n_1	12	n_2	20	

$\bar{x}_1 - \bar{x}_2 = 10 - 4 = 6$

(a) What is the distribution of the sample average difference between group 1 and group 2, if we assume that the new drug does not help people lose weight? $\mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$

$\bar{x}_1 - \bar{x}_2 \sim t\left(\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 = 0,\ SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{4^2}{12} + \frac{3^2}{20}} \approx 1.3354,\ df = \min\left\{\begin{matrix} n_1-1 \\ n_2-1 \end{matrix}\right. = n_1 - 1 = 12-1 \leq 11\right)$

(b) What is the probability that the average difference was greater than what we saw in our sample?

$$P(\bar{x}_1 - \bar{x}_2 > 6) = 1 - P(\bar{x}_1 - \bar{x}_2 < 6)$$
$$= 1 - P\left(t < \frac{6-0}{\sqrt{\frac{16}{12} + \frac{9}{20}}}\right)$$
$$= 1 - Pt\left(\frac{6-0}{\sqrt{\frac{16}{12} + \frac{9}{20}}}, 11\right) = 0.00045584 \approx 0.005$$

(c) Do you think the weight loss drug works?

Yes