Exam 3: Review

GOVT 201: Statistics for Political Science
Review Topics

Chapter 5 and 6:
1. Probability
2. Standard Deviations and Percentages of the Normal Curve
3. Table A: Z Scores
4. Z formula
5. Sampling Error (Definition)
6. Sampling Distribution of Means (Definition)
7. Standard Error (Definition)
8. Standard Error formula
9. Standard Error Unbiased formula
Probability

To employ the normal distribution to solve problems, we must acquaint ourselves with the **area under the normal curve**: the area that lies between the curve and the base line containing **100% of all the cases in any given normal distribution**.
New Symbols for Probability

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$ (mu)</td>
<td>The mean of a probability distribution</td>
</tr>
<tr>
<td>$\sigma$ (sigma)</td>
<td>The standard deviation of a probability distribution</td>
</tr>
<tr>
<td>$\sigma^2$ (sigma square)</td>
<td>The variance of a probability distribution</td>
</tr>
</tbody>
</table>

**NOTE:** Xbar, $s^2$, s represent the mean, variance and standard deviation of observed data.
Let’s take another look at standard deviation

Standard deviation is a **measure of variation**, and this variability is reflected in the sigma values (\( \sigma \)) in our distribution.

Our mean (\( \mu \)) establishes a standardized “**zero**” and our sigma values (\( \sigma \)) indicate the distance (or **variation from** \( \mu \)) of our score from the \( \mu \).
Note how the mean equals zero. Also, see how one standard deviation away from the mean is represented by $\mu + 1\sigma$ or $\mu - 1\sigma$

(depending on the direction of the deviation).
The Area Under the Curve

- $+3\sigma : 49.87\%$
- $+2\sigma : 47.72\%$
Clarifying the Standard Deviation: Men

Men:
\[ \mu = 100 \]
\[ \sigma = 15 \]
Standard Deviation: Using Table A

So far, when analyzing the normal distribution, we have looked at distances from the mean that are exact multiples of the standard deviation ($+1 \sigma$, $+2 \sigma$, $+3 \sigma$ or $-1 \sigma$, $-2 \sigma$, $-3 \sigma$). How do we determine the percentages of cases under the normal curve that fall between two scores, say $+1 \sigma$, $+2 \sigma$ for example.

Example: $\sigma=1.40$
What is the percentage of scores that fall between the mean ($\mu$) and $\sigma=1.40$. Since $\sigma=1.40$ is greater than 1, but less than 2, we know it includes more than 34.13% but less than 47.72%. 
Standard Deviation: Using Table A

\[ \sigma = 1.4 \]

-3 Standard deviations
-2 Standard deviations
-1 Standard deviation

Mean

34\% 34\%
14\% 14\%
2\% 2\%

34.13\%
47.72\%
?\%

\( \sigma = 1.4 \)
Standard Deviation: Using Table A

To determine the exact percentage between the mean (\( \mu \)) and \( \sigma = 1.40 \), we need to consult Table A in Appendix B.

Table A: Shows you the percent under the normal curve and:

- **Column A**: The sigma distances are labeled \( z \) in the left-hand column
- **Column B**: The percentage of the area under the normal curve between the mean and the various sigma distances from the mean
- **Column C**: The percentage of the area at or beyond various scores toward either tail of the distribution.
Standard Deviation: Using Table A

Using Table A:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>µ and z</td>
<td>beyond Z</td>
</tr>
<tr>
<td>1.40</td>
<td>41.92</td>
<td>8.08</td>
</tr>
</tbody>
</table>

\[ \sigma = 1.4 \]
Z Score Computed by Formula

We obtain the z score by finding the deviation \((X - \mu)\), which gives the distance of the raw score from the mean, then dividing this raw score deviation by the standard deviation.

\[
 z = \frac{X - \mu}{\sigma}
\]

*where*  
\(\mu\) = Mean of a distribution  
\(\sigma\) = standard deviation of a distribution  
z = standard score
Z Scores

Z Scores:
The \textit{z score} indicates the direction and degree that any given \textit{raw score deviates from the mean} in a distribution on a \textit{scale of sigma units}.

\[
Z = \frac{X - \mu}{\sigma}
\]
Z Scores

So why do we use z scores?
Z scores allow us to translate any raw score, regardless of unit of measure, into sigma units (standard deviation within a probability distribution) or standard deviation into a percentage on the normal curve, which provide us with a standardized/normalized way to evaluation the variation of raw scores from a standardized mean.

BUT, the sigma distance is specific to particular distributions. It changes from one distribution to another.

For this reason, we must know the standard deviation of a distribution before we are able to translate any particular raw score into units of standard deviation.
We have covered finding probability and z scores, so let’s discuss finding probability under the normal curve.

The **normal curve** can be used in conjunction with **z scores** and **Table A** to determine the probability of obtaining any raw score in a distribution.

Remember, the normal curve is a probability distribution in which the total area under the curve equals 100% probability.
The central area around the mean is where the scores occur most frequently.

The extreme portions toward the end are where the extremely high and low scores are located.

So, in probability terms, probability **decreases as we travel along the baseline away from the mean in either direction.**

To say that 68.26% of the total frequency under the normal curve falls between -1σ and +1σ from the mean is to say that the probability is approximately 68 in 100 that any given raw score will fall in this interval.
Example: Campaign Phone-Bank
We are asked to calculate the z-score for the number of calls campaign volunteers made in a 3-hour shift.

The mean number of calls is 21 with a standard deviation of 1.45σ. What is the probability that a volunteer will complete 25 calls during the 3 hour period?

Let’s apply the z-score formula.
\[ z = ? \]

\[ \mu = 21 \]

\[ \sigma = 1.45 \]

\[ X = 25 \]

\[
\begin{align*}
  z &= \frac{X - \mu}{\sigma} \\
  &= \frac{25 - 21}{1.45} \\
  &= 2.75
\end{align*}
\]

Remember our equation.

Plug in our values and scores.

We have our z score.

From our z score, we know that a raw score of 25 is located 2.75σ above the mean.
Our next step is to use Table A to find the percent of the total frequency under the curve falling between the z score and the mean.

So,

1. Let’s find our z score (2.75) in Column A.
2. Column B tells us that 49.70% of all volunteers should be able to complete between 21 and 25 calls in 3 hours.
3. By moving the decimal two places to the left, we see that the probability is 50 in 100 (rounding up).
4. Or $P = .4970$ that a volunteer will complete between 21 and 25 phone calls.
\( \sigma = 1.45 \)

\( z = 2.75 \)

49.70 or 50 in 100

50 in 100 or \( P = .4970 \)
P of Calls: 21-25

$P = 0.4970$

50 in 100

$\sigma = 1.45$

$z = 2.75$

49.70 or 50 in 100

$P = 0.4970$

0.30

21

25

Mean
P of Calls: 17-25
P = .9940
99 in 100

P of Calls: less 17, more 25
P = .0060
.6 in 100
Samples and Populations

Statistics for Political Science
Levin and Fox
Chapter 6
Sampling Error

Sampling Error:
With random sampling, we can always expect some difference between a sample (random or nonrandom) and the population from which it was drawn.

This is known as sampling error. Sampling error is in every study no matter how rigorous the design or careful the researcher. As such, the sample mean ($\bar{X}$) will never be exactly the same as a population mean ($\mu$).
Sampling Error

Sampling Error: Table 6.1

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample A</th>
<th>Sample B</th>
<th>Sample C</th>
</tr>
</thead>
<tbody>
<tr>
<td>70 80 93</td>
<td>96</td>
<td>40</td>
<td>72</td>
</tr>
<tr>
<td>86 85 90</td>
<td>99</td>
<td>86</td>
<td>96</td>
</tr>
<tr>
<td>56 52 67</td>
<td>56</td>
<td>56</td>
<td>49</td>
</tr>
<tr>
<td>40 78 57</td>
<td>52</td>
<td>67</td>
<td>56</td>
</tr>
<tr>
<td>89 49 48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99 72 30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96 94</td>
<td>$\bar{X} = 75.55$</td>
<td>$\bar{X} = 62.25$</td>
<td>$\bar{X} = 68.25$</td>
</tr>
<tr>
<td>( \mu = 71.55 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sampling Distribution of Means

**Sampling Distribution of Means:**

Even with *sampling error*, it is possible to generalize from a *sample to a larger population*.

What we can do is construct a *sampling distribution of means*. We can collect data on *multiple samples and calculate their means*. These means are then placed in a frequency distribution.
Consider these statements to be truths that hold for all sample distribution of means:

1. The **sampling distribution of means approximates the normal curve.** This is true regardless of the shape of the distribution of raw scores in the population from which the means are drawn, so long as the sample of means is large enough (over 30).

2. The **mean of a sample distribution of means (the mean of means)** is equal to the **true population mean.**

3. The **standard deviation** of a sample distribution of means is **smaller** than the standard deviation of the population.
Pointers for distinguishing between:

(1) the standard deviation (\(\sigma\)) of raw scores in the population and

(2) the standard deviation of the sampling distribution of sample means

Notation:

\[
\sigma_{\bar{X}} = \text{standard deviation of the sample distribution of means.}
\]

The \(\sigma\) indicates that this is an unobserved probability distribution.

The \(\sigma_{\bar{X}}\) indicates that this is the standard deviation among all possible sample means.
Sampling Distribution of Means as a Probability Distribution

- Mean of means
- Stand. dev of sample means
- 34.13%
- 47.72%
- 49.87%
- $\mu$
- $+1\sigma_x$
- $+2\sigma_x$
- $+3\sigma_x$
Standard Error of the Mean

In actual practice, the researcher rarely collects data on more than one or two samples.

As a result, the researcher does not know the mean of means or the standard deviation of the sampling distribution. They have not collected several samples from which a series of means could be be calculated.

**Standard error of the mean**

However, the standard deviation in a theoretical (the distribution that would exist in theory if the means of all possible samples were obtained) sampling distribution can be derived.

This derivation is known as the standard error of the mean.
Standard Error of the Mean

The standard error of the mean is obtained by dividing the population standard deviation ($\sigma$) by the square root of the sample size.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Where

$\sigma$ = population standard deviation

$\sqrt{n}$ = square root of sample size
The Normal Curve: Theoretical Overview

- Measures of mean and SD
- Percentages
- Sample or pop/prob. mean
- Sample or pop/prob. SD/SE

- $X\bar{\text{bar}}/\mu$
- $+1s/\sigma/\sigma_x$
- $+2s/\sigma/\sigma_x$
- $+3/\sigma/\sigma_x$

- 34.13%
- 47.72%
- 49.87%

- Sample or pop/prob. mean
- Sample or pop/prob. SD/SE
CONFIDENCE INTERVALS

NOTE: Confidence intervals can be constructed for any level of probability.

Most researchers are not confident enough to estimate a population mean knowing that there are only 68 out of 100 chances of being correct.

So, we generally use a wider, less precise confidence interval having a better probability.

The standard is 95% confidence interval, when we accept that there is a 5% chance that we could be wrong!
Confidence Intervals

So how do we find a 95% confidence interval?

1. We look at our normal distribution and know that 95.44% of the sample means lies between $-2 \sigma \overline{X}$ and $+2 \sigma \overline{X}$ from the mean of the means.

2. If we divide 95 by 2 (for both sides of the standard deviation from the means), we get 47.5.

3. Go to Table A and look in column b for 47.5 which gives us a z score of 1.96.

$$95\% \text{ confidence interval} = \overline{X} \pm 1.96 \sigma \overline{X}$$

If we apply this to our estimated mean from our sample, we would plug the values into the equation above and multiply 1.96 by the value we calculated for $\sigma \overline{X}$. 

37
The t Distribution

At this point, we want to make statements about our population based on information that we have about our sample.

We know that our sample mean and sample distribution will be smaller than those for our population. That is, in any given sample, the sample variance ($s^2$) and the sample standard deviation ($s$) calculated from a sample mean will be smaller than if calculated from the population mean ($\mu$).

Specifically, sample variance and sample standard deviation tend to be smaller than population variance and the sample standard deviation.
The t Distribution

Since the **sample estimates are biased**, we will need to **inflate** them to produce more accurate estimates of the **population variance** and **population standard deviation**.

We do this by dividing by **$N - 1$** rather than just **$N$** when calculating **variance** and **standard deviation**.
Unbiased Estimates of the Population Variance and Population Standard Deviation

Population Variance:

\[ \hat{\sigma}^2 = \frac{\sum (X - \bar{X})^2}{N - 1} \]

Population Standard Deviation:

\[ \hat{\sigma} = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}} \]
An unbiased estimate of the standard error of the mean is:

\[ s_{\bar{x}} = \frac{s}{\sqrt{N - 1}} \]

Where \( s \) is the sample standard deviation from a distribution of raw scores or from a frequency distribution.
The t Distribution

There is an complication when it comes to using an unbiased estimate of the standard error of the mean when calculating z scores in a normal distribution.

When we estimate the standard error using sampling data rather than population data, the sampling distribution of means is a bit wider (more dispersed).

That is when we use:  
\[ z = \frac{\bar{X} - \mu}{s_X} \]

Instead of:  
\[ z = \frac{\bar{X} - \mu}{\sigma_X} \]
So, the ratio does not allow us to follow the $z$ or normal distribution.

We will need to follow a t distribution, rather than a z distribution, which will give us a t ratio.

$$t = \frac{\bar{X} - \mu}{s/\sqrt{X}}$$
The t Distribution

The degrees of freedom indicate how close the t distribution comes to approximating the normal curve. It is a means of indicating the difference between a sample and a probability distribution.

When estimating the population mean, the degrees of freedom are one less than the sample size.

\[ \text{df} = N - 1 \]

Where \( \text{df} \) = degrees of freedom

\( N \) = sample size
The **greater** the degrees of freedom, the **larger** the sample size, the **closer** the **t distribution** gets to the normal distribution. That is, the reliability of our estimate of the standard error of the mean increases as our sample increases, so **t** approaches **z**.

Rather, the **reliability** of our estimate of the standard error of the mean **increases** as our sample size increases, and so the **t ratio** approaches a **z score**.
When working with the t distribution, we use Table C instead of Table A.

Table C is calibrated for various levels of alpha (\(\alpha\)).

**Alpha** is the area in the tails of the distributions and the formula for calculating alpha is

\[
\alpha = 1 - \text{level of confidence}
\]

95% level of confidence, \(\alpha = .05\)  
99% level of confidence, \(\alpha = .01\)

To use Table C, we need two pieces of information:

1. The degrees of freedom \((N - 1)\)
2. The alpha value \((1 - \text{level of confidence})\)
The larger our sample, the larger our df, and the smaller our t values.

The larger our df and the smaller the t values, the closer they approach the z or normal distribution.

SO, when the standard error of the mean is estimated, we can construct a confidence interval using the t values from Table C.

\[ \text{CI} = \overline{X} \pm t s \frac{\overline{X}}{X} \]

NOTE: You multiply the t value by the standard error of the mean.

ALSO: t multiplied by s \( \frac{\overline{X}}{X} \) provides us with our margin of error.
Sample and Populations

\[ \bar{X} : \text{Mean of Sample} \]
\[ \mu : \text{Mean of Population/Prob} \]
\[ s : \text{Stand. Dev. of Sample} \]
\[ \sigma : \text{Stand Dev of Population/Prob} \]

Sampling Error

-Difference between Sample and Population

Sample Distribution of Means (SDM)

-Distribution of sample means
-Approx Normal Curve
-Means of Means = True pop mean
-SD of SDM is smaller than SD or pop

SDM and Normal Curve

-Can use Normal Cur. to make P statements about sample means
-Can use z scores

Standard Error

\[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \]  
Pop. stand. dev divided by the square root of the sample size.

Confidence Intervals

\[ 68\% = \bar{X} \pm \sigma_{\bar{X}} \]
\[ 95\% = \bar{X} \pm 1.96\sigma_{\bar{X}} \]

Details:

Formulas:

\[ \bar{X} - \mu \]
Sample mean, minus pop. mean, divided by the stand dev. of the sampling distribution of means

\[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \]  
standard deviation of the sampling distribution of means.

\[ z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \]  
Est. range of true pop. mean.
Details:
-Used when we want to make statements about pop. from samples

The t distribution

Sample Variance and Stand. Dev.: biased
- Sample var. ($s^2$) and sample stand. dev. ($s$) is smaller than pop. var. and the pop. standard deviation.

Pop. Variance and Pop. Stand. Dev.: unbiased
- To "inflate" sample var. ($s^2$) and sample stand. dev. ($s$), we use $N - 1$, rather than just $N$.

Standard Error: Unbiased
- Used to calculate an unbiased standard error.

Formulas:
- Sample variance and stand. dev.: $\hat{\sigma}^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$ vs. $\sigma^2 = \frac{\sum (X - \bar{X})^2}{\sqrt{N - 1}}$

- Standard error: $s_{\bar{X}} = \frac{s}{\sqrt{N - 1}}$

- t rather than z score
  $t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$

- df = $N - 1$

- $\alpha = 1 - \text{level of confidence}$

- 95/99% level of confidence, $\alpha = .05/.01$
Review: Samples and Populations

Type of Data: Standard Error (lack SDM: theoretical population data)
Population Standard Deviation: $\sigma$
Population Mean: $\mu$
Sample data: $N=?$

Formula:

95% Confidence Level:
Review: Samples and Populations

Type of Data: Standard Error: Unbiased (No SDM, only sample data)

Sample Standard Deviation: \( s \)

Sample Mean: \( \bar{X} \)

Sample data: \( N \)

1. Formula: \( s_{\bar{X}} = \frac{s}{\sqrt{N-1}} \)

2. Degrees of freedom: \( df = N - 1 \)

3. Alpha: \( \alpha: 1 - \text{level of confidence} \) (95%: 1 - .95 = .05)

4. Consult Table C to get t score (based on \( df \) and \( \alpha \))

4. 95% Confidence Level: \( = \bar{X} \pm ts_{\bar{X}} \)
Extra: Sample and Populations

Sample and Populations: Symbols
Now we have to carefully distinguish between samples and populations.

**Mean:**
\[ \overline{X} = \text{mean of a sample} \]
\[ \mu = (\text{mu}) \text{ mean of a population/probability} \]

**Standard Deviation:**
\[ s = \text{standard deviation of a sample} \]
\[ \sigma = (\text{sigma}) \text{ standard deviation of a population/probability} \]

**Note:** The symbols for population and probability distributions are the same.
Extra: Sampling Error

Sampling Error:
With random sampling, we can always expect some difference between a sample (random or nonrandom) and the population from which it was drawn.

This is known as sampling error. Sampling error is in every study no matter how rigorous the design or careful the researcher. As such, the sample mean ($\bar{X}$) will never be exactly the same as a population mean ($\mu$).