Sampling Distributions & One Sample Tests Notes[1]

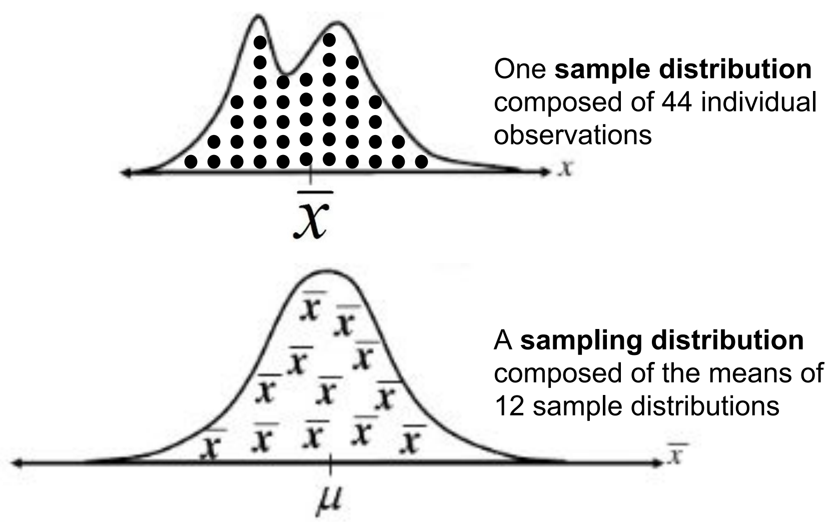
Stanford GSE Math Camp

Do Not Distribute Outside GSE

1 Sample Distributions and Sampling Distributions

Sampling Distributions are a critical, foundational understanding in statistics, but many confuse them with sample distributions. What is the difference?

- **Sample Distributions** are distributions built from a sample you've taken. These are the types of distributions we've been working with so far. You have a set of data and you plot it - that's a sample distribution. Each point on the dotplot stands for an individual observation.
- **Sampling Distributions** are distributions that allow us to see how statistics describing samples (like $\bar{x}$ and s) vary from sample to sample. Think of a sampling distribution as a distribution of many *sample distributions*, or at least some statistics describing these sample distributions.
  - Example Imagine you roll a die 10 times. After the ten rolls, you plot the values for all the rolls on a distribution. This is a *sample distribution.* Then, you calculate the mean of this sample distribution. This mean gets placed, as an observation, on a new distribution. You repeat this process several times. This second distribution, where you place the calculated means, is called the sampling distribution.



2 Sampling Distributions for Proportions

- What is a proportion? A proportion is just like a fraction, where X is the number of successes and n is the total number of trials. Proportions are **categorical data**.

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

Note: $\hat{p}$ is the notation for a sample or observed proportion

[1] Contributor(s): Kelly Boles & Erin Fahle. If you find errors, please let us know so that we may correct them. Thanks!

- Just like means, proportions have symbols and formulas for both their means and standard deviations.

	Mean	Standard Deviation
Population Proportion	p	$\sqrt{\frac{pq}{n}}$

Note: p is the probability of success. q is the probability of failure. n is the number of trials.

- We calculate z-scores for proportions very similarly to z-scores for means. $z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$
- The mean for the sampling distribution for proportions is given by:

$$\mu_{\hat{p}} = p$$

- The standard deviation for the sampling distribution for proportions is given by:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

- We must check three conditions in order to use a normal model for a *sampling distribution for proportions*:
  - As long as we meet the **Randomization Condition** (random sample)
  - **Success/Failure Condition**, which says that the expected number of successes (n*p) and the expected number of failures (n*q) must be greater than or equal to 10, and
  - **10% Condition**, which says that the sample size, n, must be no larger than 10% of the population,

  ...then the sampling distribution for proportions can be described as:

$$\sim \text{N}(\text{p}, \sqrt{\frac{pq}{n}})$$

3 1 Sample Z Test for Proportions (See handout)

To conduct a 1 Sample Z Test for Proportions, we must follow these steps:

1. Name test, write hypothesis & describe parameter
2. Check assumptions/conditions
   - Independence Assumption
   - Randomization Condition
   - Success/Failure Condition
   - 10% Condition
3. Do calculations to get test statistic & p-value
   - Essentially, what a test statistic is asking is: ASSUMING that your distribution did have a mean of p, how likely is it that you observed the sample mean, $\hat{p}$? If the test statistic is really big (in absolute value), it is really unlikely that it could have been observed if the true mean was really p.

4. Interpret results & draw conclusions
    - If the p-value is smaller than some *α-level* or *significance level*, we will reject the null hypothesis and say the mean is *statistically significantly* different/greater than/less than from the null hypothesis.
    - Typically, we use $\alpha = 0.001, 0.01, 0.05, 0.10$.

4 Sampling Distributions for Means

The Sampling Distribution for Means is similar to that of proportions but a bit more complex.

- The mean for the sampling distribution of means is given by:

$$\mu_{\bar{x}} = \mu$$

- The standard deviation for the sampling distribution of means is given by:

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

  - *Note: This formula requires that we know the population standard deviation. If we are trying to make predictions about the population mean, it's incredibly unlikely we'll know the population standard deviation.*

- If we don't know the population standard deviation, then we need to estimate it. When we are forced to estimate a parameter, the calculation is called a **standard error** rather than a standard deviation. The formula for **standard error of the mean (SEM)** is:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Note: We are using the sample standard deviation in the numerator to estimate the population standard deviation.

- Just like in the sampling distribution for proportions, we must check conditions. Most conditions are the same.
  - **Randomization Condition**(random sample)
  - **10% Condition**, which says that the sample size n should be no more than 10% of the population
  - **Nearly Normal Condition**, which can be met if (1) the population is normal, (2) you are given a dataset and its graph is normal, or (3) the Large Enough Sample Condition is satisfied
    * **Large Enough Sample Condition**, which states that, because of the Central Limit Theorem, large enough samples will result in a sampling distribution that fits a normal model. There is no rule for how large a sample you'll need, but a typical rule of thumb is n$\geq$30. There are also other ways to check for normality, if your sample is small.
- Thus,if conditions are met, the sampling distribution for means is given by:

Known σ	Unknown σ
$\sim$N($\mu, \frac{\sigma}{\sqrt{n}}$)	$\sim$N($\mu, \frac{s}{\sqrt{n}}$)

5 1 Sample Z or T Test for Means (See handouts)

To conduct a 1 Sample Z or T Test for Means, we must follow these steps:

1. Name test, write hypothesis & describe parameter
2. Check assumptions/conditions
   - Independence Assumption
   - Randomization Condition
   - Nearly Normal Condition
   - 10% Condition
3. Do calculations to get test statistic & p-value.
   - We'll have to determine whether a z-test or t-test is appropriate. We use a z-test when the population standard deviation is known (σ) or the sample size is large. Otherwise, t-tests work best for small samples.
     - While you calculate a t-score just like you would a z-score (but using a t table), the t distribution has an additional parameter called degrees of freedom. For one sample tests, degrees of freedom are given by n-1.
4. Interpret results & draw conclusions
   - If the p-value is smaller than some *α-level* or *significance level*, we will reject the null hypothesis and say the mean is *statistically significantly* different/greater than/less than from the null hypothesis.
   - Typically, we use $\alpha = 0.001, 0.01, 0.05, 0.10$.