Lecture: Sampling Distributions and Statistical Inference

Sampling Distributions

population – the set of all elements of interest in a particular study.

sample – a sample is a subset of the population.

random sample (finite population) – a simple random sample of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

random sample (infinite population) – A simple random sample from an infinite population is a sample selected such that each element selected comes from the same population and each element is selected independently.

Point Estimation

At some point we may be interested in taking a random sample from a population and estimating population parameters. In particular there are two parameters that we will be estimating a lot.

The sample mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is an estimator of the population mean μ.

The sample variance $s^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}$ is an estimator of the population variance σ^2.

What we eventually want to be able to do is

(1) Compute the sample mean and variance.
(2) Use this sample mean and variance to make inferences and test hypothesis about the population mean.

Example: Given that the mean height is 69 inches how likely is it that the sample mean is more than 72 inches?

Right now we do not know how to answer questions like this. Today we are going to develop a tool that is going to be very useful in helping us do statistical inference. This tool is the Central Limit Theorem.

Expectation and Variance of the Sample Mean

Before I give you the central limit theorem I want to go through a few things.

Expectation of the sample mean: $E(\bar{x}) = E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n}(n \cdot \mu) = \mu$

We have already computed the mean of the sample mean. Lets see if we can compute the variance of the sample mean.

$$Var(\bar{x}) = \frac{\sum_{i=1}^{n} Var(x_i)}{n^2} = \frac{\sigma^2}{n}$$

Now we know the mean and the variance of $\bar{x}$. The only thing that we do not know is how $\bar{x}$ is distributed. The most important theorem is statistics tells us the distribution of $\bar{x}$.

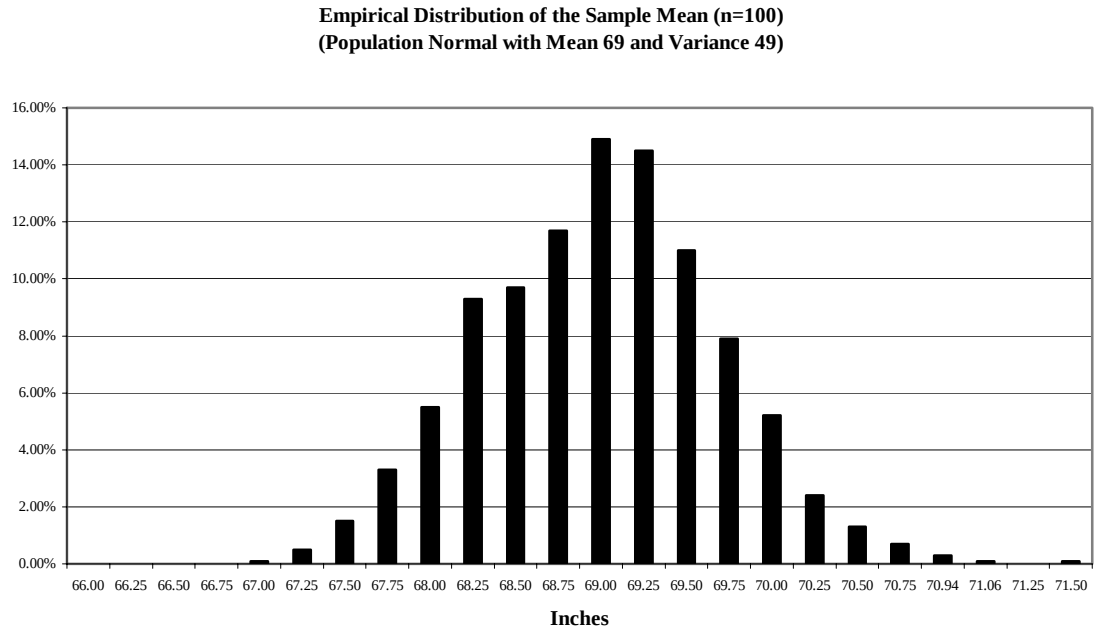
Central Limit Theorem: In selecting a sample size n from a population, the sampling distribution of the sample mean can be approximated by the normal distribution as the sample size becomes large.

In particular if the population is infinite (or very large) $\frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{\sqrt{n}\left(\bar{x} - \mu\right)}{\sigma} \sim N(0,1)$

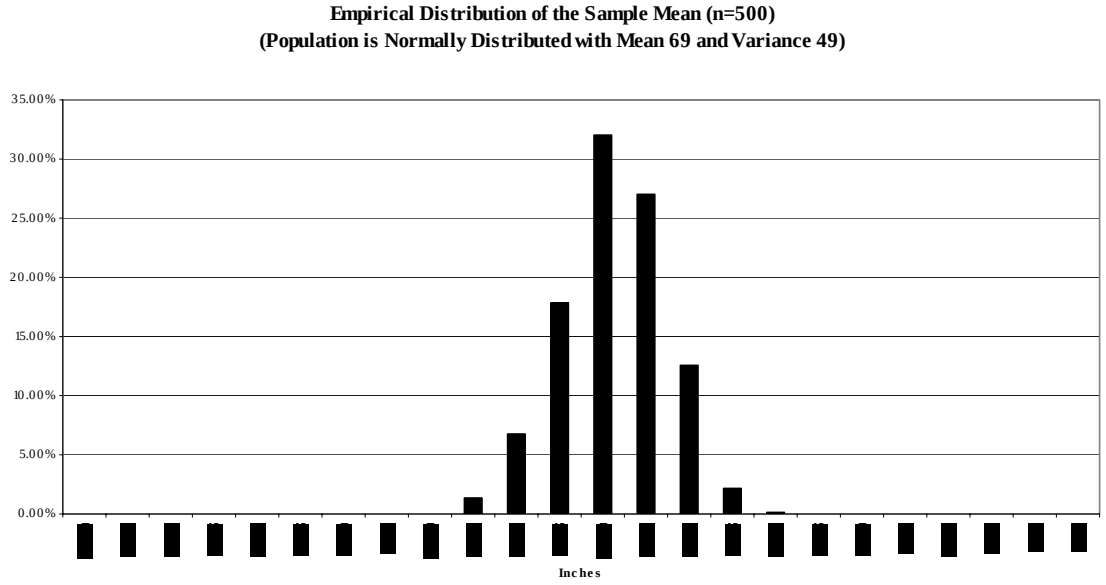
$\Rightarrow \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Example: Suppose that (in the population) human height is distributed normally with mean 69 and variance 49. What is the distribution of the sample mean? It will depend on the how large my sample is.

Sample Size=100

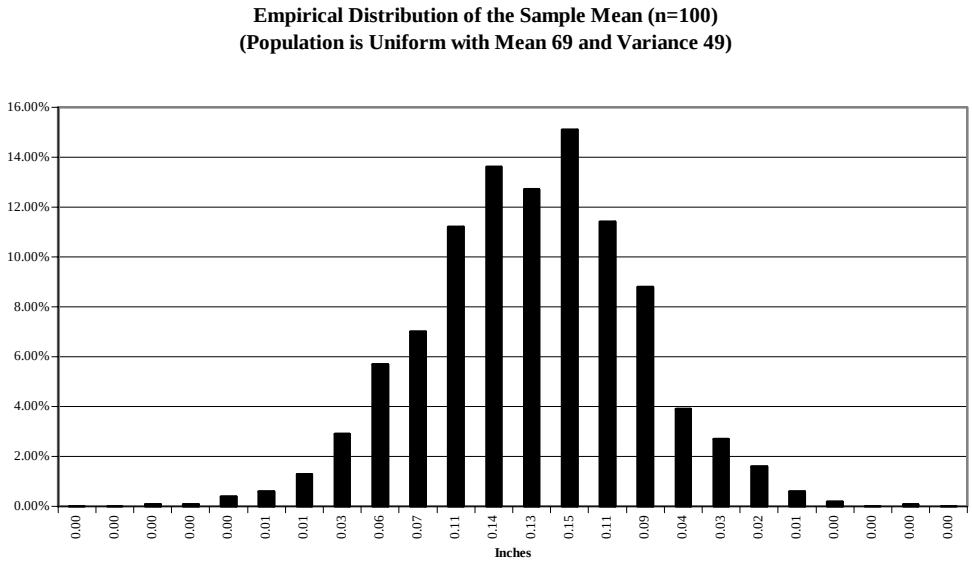


Sample size=500

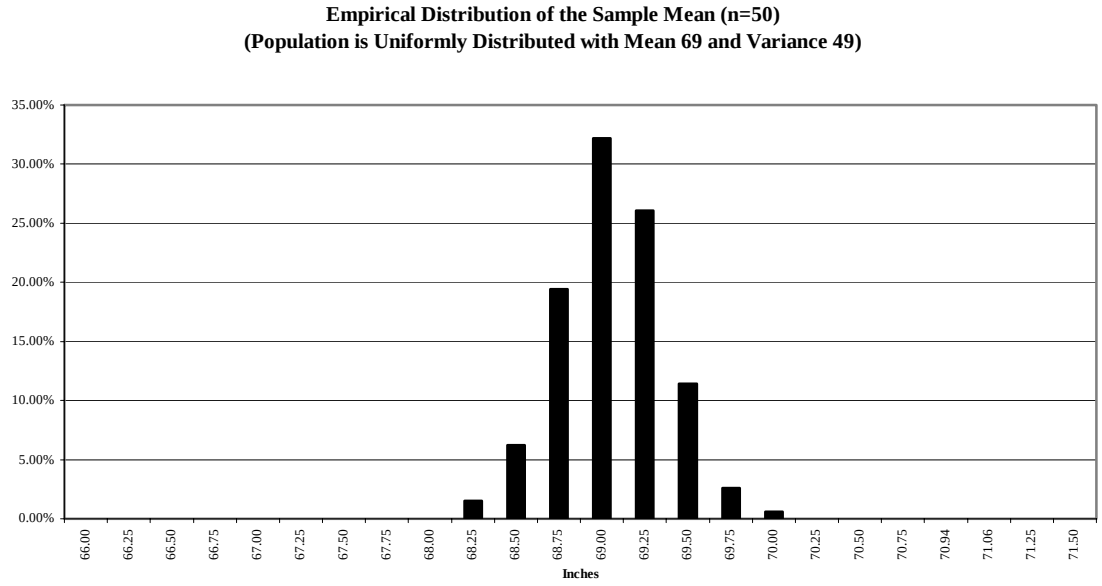


Example: Suppose that (in the population) human height is distributed uniformly with mean 69 and variance 49. What is the distribution of the sample mean? It will depend on the how large my sample is.

Sample Size=100



Sample Size=500



Very Important – The distribution of the mean does not dependent on the underlying distribution of the variable in question.

The Sampling Distribution of $\bar{x}$

We are able to show

$$E(\bar{x}) = \mu \text{ and } Var(\bar{x}) = \frac{\sigma^2}{n}.$$

The Central Limit Theorem also tells us that the distribution of $\bar{x}$ can be approximated by the Normal Distribution if the sample size is large. Putting this information together with what we know about the mean and variance of the sample average we get

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

This is a very powerful result as it indicates that regardless of the distribution of the x's, the distribution of the sample average is approximately normal if the sample size is big enough.

Suppose that I am interested in the income of working age men. I go out and take a sample of 10,000 men and find that the sample average of income is $30,000 and that sample standard deviation of income is $20,000.

The Central Limit theorem is useful as it allows us to make inferences about the sample mean.

I have a hypothesis that average income among working age men is $40,000. Is it possible that me hypothesis is correct given that I obtained a sample mean of $30,000 and a sample standard deviation of 20,000. Using the tools we just developed I can answer this question. How?

(1) I assume that my hypothesis is correct. That is I assume that $\mu = \$40,000$.

(2) I form a Z statistic where I use s^2 as an estimate of σ^2. This looks something like

$$z = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{\sqrt{n}\left(x - \mu\right)}{s} = \frac{\sqrt{10,000}\left(30,000 - 40,000\right)}{20,000} = \frac{-100 \cdot 10,000}{20,000} = -50$$

Some Terminology

Sampling Error or Error - $\bar{x} - \mu$ is often referred to as the sampling error of just the error. It is the error that you make when attempting to estimate μ by $\bar{x}$.

Standard Error ($\sigma_{\bar{x}}$)– The standard error is simply the standard deviation of the sampling error. The standard deviation of the sampling error is simply the standard deviation of $\bar{x}$ because μ is a constant - $\frac{\sigma}{\sqrt{n}}$

Interval Estimation

Given a random sample we can estimate the population mean with the sample mean. What I have tried to show you is that the sample mean can be quite variable with the degree of variability depending on the sample size and the population variance $Var(\bar{x}) = \frac{\sigma^2}{n}$.

Much of the time we may want to use the sample mean to estimate an interval which will contain the population mean with a specific level of confidence or we may want to include confidence bounds in our estimate of the population mean.

When you get the results of a poll you often see (in small percent) plus or minus 5 percent (or something like this). The plus or minus 5 percent are confidence bounds. We will learn how to find them today.

To do this we are going to have to use what we know about the location of the sample mean with respect to the population mean.

We know that the population mean is within 1 standard deviation of the sample mean 68 percent of the time

We know that the population mean is within 1.645 standard deviations of the sample mean 90 percent of the time.

We know that the population mean is within 1.96 standard deviations of the sample mean 95 percent of the time.

Large Sample Case (σ^2 unknown)

Example: Lets take the information that I gave you from the first problem last week and estimate a 95 percent confidence interval

$$\left.\begin{array}{l}\bar{x} = 30,000 \\ s = 20,000\end{array}\right\} \Rightarrow Var(\bar{x}) = \frac{\sigma^2}{n} \approx \frac{s^2}{n} = 40,000 \Rightarrow \sigma_{\bar{x}} = \text{standard error} = 200$$

We know that

$$P(-1.96 \le \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \le 1.96) = 0.95 = P(-1.96 \cdot \sigma_{\bar{x}} \le \bar{x} - \mu \le 1.96 \cdot \sigma_{\bar{x}}) \Rightarrow$$

$$P(-1.96 \cdot \sigma_{\bar{x}} + \bar{x} \le -\mu \le 1.96 \cdot \sigma_{\bar{x}} + \bar{x}) = P(-1.96 \cdot \sigma_{\bar{x}} - \bar{x} \le -\mu \le 1.96 \cdot \sigma_{\bar{x}} - \bar{x}) =$$

$$P(\bar{x} - 1.96 \cdot \sigma_{\bar{x}} \le \mu \le \bar{x} + 1.96 \cdot \sigma_{\bar{x}})$$

So we can be 95% confident that that the population mean is between 29,608 and 30,392 (because $1.96 \cdot \sigma_{\bar{x}} = 392$)

If we me a statement like the sample mean income among working age men in the population is between 29,500 and 30,500 with 95 percent confidence.

Here 95% is the **confidence level**.

The **confidence coefficient** is 0.95.

The **significance level** (α) = 1 – Confidence Coefficient

Formulas (Again you do not really need these of understand the intuition behind them)

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = z_{\alpha/2} \cdot \sigma_{\bar{x}}$$

where s is the sample standard deviation (use the population standard deviation if you have it, $1-\alpha$ is the confidence coefficient, and $z_{\alpha/2}$ is the z value providing an area $\alpha/2$ in the upper tail of a standard normal probability distribution.

Example: Suppose that we wanted a 90% confidence interval. We just use different z values in your calculations. Here we want $z_{0.05}$ instead of $z_{0.025}$

Lower bound = 30,000-1.645*200=29,669.2
Upper bound=30,000+1.645*200=30330.8

We can be 90% confident that the population mean is between 29,669.2 and 30330.8.

Small Sample Case

Here we do not really have the central limit theorem because the central limit theorem relies on the sample size being large.

Basically, in order to test hypothesis or estimate confidence bounds we have to make the rather dubious assumption that the population is normal.

If the sample is small and we know the population variance (hardly ever). Then we can use the formula supplied above.

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

If the population standard deviation is not known then we use

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = t_{\alpha/2} \cdot \sigma_{\bar{x}}$$

So instead of using the standard normal table we will use the table associated with the students t distribution with n-1 degrees of freedom. The t distribution is like the normal (symmetric and bell shaped), but is has thicker tails. As the degrees of freedom (read sample size) gets bigger the student t distribution looks more and more like a normal distribution.

Note: Technically speaking we are always using the t-distribution when the population variance σ^2 is unknown. Its just that when the sample is large there is no discernable difference between the t- and normal distributions.

Note that normal tables give you the CDF evaluated a given value, the t tables give you the t that leave 0.10, 0.05, 0.25, 0.01, and 0.005 in the upper tail for different degrees of freedom.

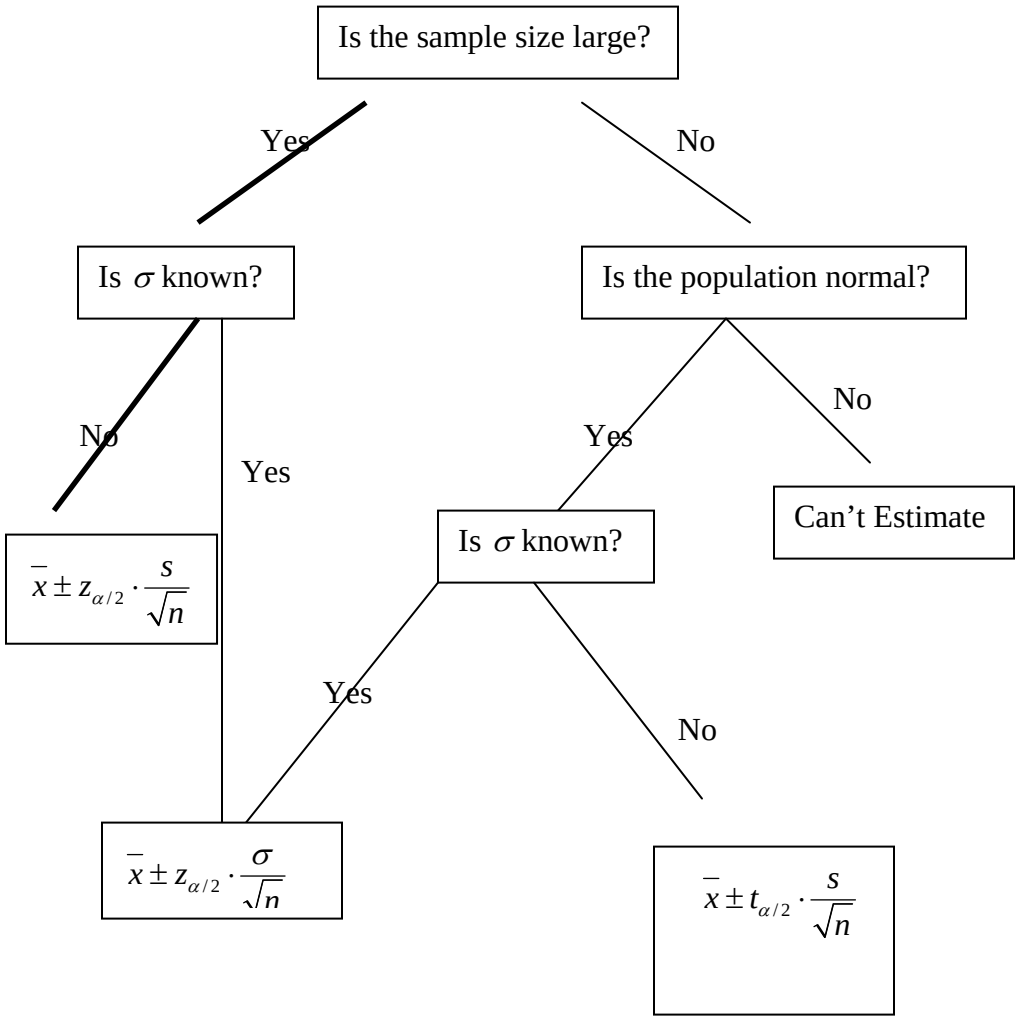
Example: Suppose I took a sample of 16 working age men and found an average income of 30,000 and a sample standard deviation of 20,000. I want to be able to estimate a 90 percent confidence interval.

Note that $\sigma_{\bar{x}} = \frac{20,000}{\sqrt{16}} = 5000$

Upper Bound: 30,000+1.753*5000=38,765
Lower Bound: 30,000-1.753*5000=21,235

We can be 90% confident that mean income is between 21,235 and 38,765 assuming that income among working age men is normally distributed (which it is not). Note that we have a much wider confidence band with a sample of 16 than with 10,000 and we are forced to make assumptions about the distribution of the population.

When to use the t- and normal distributions to estimate confidence intervals



Calculating n to produce a given confidence bounds

I am going to take a sample of working age men and I want to use this sample to estimate income. I want to be able to construct a 90% confidence interval that covers a range of 1000. How large should my sample be assuming the population standard deviation is 20,000?

Basically I want

$$1.645 \cdot \frac{20,000}{\sqrt{n}} = 1000 \Rightarrow$$
$$\sqrt{n} = 1.645 \cdot \frac{20,000}{1,000} \Rightarrow$$
$$n = \left(1.645 \cdot 20\right)^2 \approx 1083$$

There is another formula here, but I think you are would be better off not trying to remember it

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

where E is the desired margin of error.

Confidence Intervals for Sample Proportions

We know that if n is large

$$\bar{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

I take a sample of 500 and find that $\bar{p} = 0.40$. I want to construct a 95% confidence interval.

Note: That for confidence intervals involving sample proportions there are not large and small sample cases. Because the population is inherently not-normal, confidence intervals can only be estimated this way in cases where the sample size is large.

Hypothesis Testing

The Null and Alternative Hypothesis

The first thing that we do in hypothesis testing is make a tentative assumption about a population parameter (at this point μ).

null hypothesis (H_0) – the tentative assumption that is made about a population parameter.

alternative hypothesis (H_a) – to converse of the null hypothesis.

Example: I take a sample (income) and find that $\bar{x}$=35,000. I want to test the hypothesis that mean income is greater than 37,000. Then

$$H_0: \ \mu > 37,000$$
$$H_a: \ \mu \leq 37,000$$

After I conduct the test I will either conclude that the null hypothesis can or cannot be rejected. If the null can be rejected then I will accept the alternative hypothesis.

There are basically three combinations of null and alternative hypothesis that we will be interested in.

$$H_0: \ \mu > 37,000$$
$$H_a: \ \mu \leq 37,000 \qquad (1)$$

$$H_0: \ \mu < 37,000$$
$$H_a: \ \mu \geq 37,000 \qquad (2)$$

$$H_0: \ \mu = 37,000$$
$$H_a: \ \mu \neq 37,000 \qquad (3)$$

Type I Versus Type Two Errors

When testing hypothesis there are essentially two errors that can be made

1) Accepting the null hypothesis when the alternative is correct – Type II error
2) Accepting the alternative hypothesis when the null is correct – Rejecting a true null - Type I error.

significance level – the probability of a Type I error.

When we conduct hypothesis test we will always have to specify a significance level. We basically have to answer the question – "What is the probability of a Type I error we are willing to live with?"

We cannot generally quantify the probability of a type II error. To avoid making Type II errors we will always say we "Cannot Reject the Null Hypothesis" instead of we "Accept the Null Hypothesis."

One more probability that we will be concerned with is the p-value. The p-value is the probability of obtaining a sample result that is at least as likely as what is observed. Sounds confusing – but it really is not.

Steps in Hypothesis Testing

1) Develop the null and alternative hypothesis.
2) Specify a level of significance α
3) Select the tests statistic that will be used the test the hypothesis (for use right now this is the Z statistic).
4) Collect the sample data and compute the value of the test statistic.
5) Use the value of the test statistic to compute the p-value
6) Reject H_0 if p-value$< \alpha$

One Tailed Tests

$$H_0: \mu > \mu_0$$
$$H_a: \mu \leq \mu_0 \quad (1)$$

$$H_0: \mu < \mu_0$$
$$H_a: \mu \geq \mu_0 \quad (2)$$

Example: I believe mean income among working age men is greater than 35,000.

1) I want to test the hypothesis

$$H_0: \mu \geq 35,000$$
$$H_1: \mu < 35,000$$

2) I need to pick a significance level – lets go with 0.05.
3) I am going to use a Z statistic with s as an estimate of σ.
4) I go out and take a sample of 10,000 working age men and ask them about their income last year. I obtain $\bar{x} = 33,000$ and $s = 25,000$. The value of my test statistic is

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{33,000 - 35,000}{\left(\frac{25,000}{100}\right)} = \frac{-2,000}{25,000} = \frac{-2}{25} = -0.08$$

5) Compute the p-value

$$p - value = P(Z \leq -0.08) = 0.4681$$

6) We cannot reject the null hypothesis that $\mu \geq 35,000$ because the p-value is greater than the significance level. The probability that I would obtain a sample mean of 33,000 or lower if the true population mean was 35 is 0.4681 which is greater than the level of significance.

Another way to solve these problems is to find critical values. The criticle value is the value that leaves the level of significance to the in the tail corresponding to the null hypothesis. In the case of this example our critical value would be –1.645. Any Z value smaller than –1.645 and we would reject the null.

Two Tailed Tests

$$H_0: \ \mu = \mu_0$$
$$H_a: \ \mu \neq \mu_0$$

Basically these problems will be really easy for us if we understand interval estimation. If we to conduct a two tailed test with level of significance α, then form a confidence interval with confidence level $1-\alpha$. If the interval includes μ_0 then the null hypothesis cannot be rejected. If the interval does not include μ_0 then we should reject the null hypothesis.

Our critical values are the same s the Z values when we where estimating intervals.

What Distribution to Use?

- If the sample size is large the z-distribution may be used to get critical values or to compute p-values.
- If the sample size is small, the population is normal, and σ^2 is known the z-distribution may also be used.
- If the sample size is small, the population is normal, and σ^2 is estimated with s^2 the t-distribution must be used to calculate p-values and/or compute critical values.

Note: Whenever the hypothesis involves a population proportion you must have a large sample or use different techniques to conduct the hypothesis test because the population is inherently non-normal (its binomial) (see the Hannity and Colmes problem from Homework #2) .

Example: I intended to write an exam with a mean score of 50. Lets test the hypothesis that the mean score (among a population of public affairs students in which you are a random sample) is greater than or equal to 50 with a significance level of 0.10.

Compute the z state.

$$\frac{42.22-50}{\left(\frac{17.48}{7}\right)} = \frac{-7.68}{2.5} = -3.07$$

Find the p-value

$$p-value = P\left(\begin{matrix}\mu = 50 \\ \bar{x} \le 42.22 \mid s = 17.48 \\ n - 49\end{matrix}\right) = P(Z \le -3.07) = 0.0011$$

So we can reject the null hypothesis that $\mu \ge 50$.

Small Sample Hypothesis Testing

Question 4 from the exam with the sample size halved and a significance level of 0.05.

$\bar{x} = 9$
$s = 2$
$n = 24$

Lets test the hypothesis that $\mu \ge 10$

$H_0 : \ \mu \ge 10$
$H_1 : \ \mu < 10$

First of all we need to be able to assume that the distribution of zinc levels in this type of fish is normal.

If we can make this assumption then we may proceed.

Find our t value

$$\frac{\bar{x} - \mu_o}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{9-10}{\left(\frac{9}{\sqrt{24}}\right)} = \frac{-1}{\left(\frac{3}{\sqrt{8}}\right)} = -\frac{8}{3} \approx -2.66 =$$

Compute p-value

$$p-value = P\left(\begin{matrix}\mu = 10 \\ \bar{x} \le 9 \mid s = 2 \\ n = 24\end{matrix}\right) = P\left(t \le -2.66\right) < 0.05 = \text{significance level}$$

So we can reject the null hypothesis.

The Sampling Distribution of the Difference of Population Means.

Suppose that you have two populations – population 1 and population 2. Let μ_1 be the mean of population 1 and μ_2 be the mean of population 2.

$\bar{x}_1$ would be a point estimate of μ_1 and $\bar{x}_2$ would be a point estimate of μ_2. The note that we can take $\bar{x}_1 - \bar{x}_2$ as a point estimate of $\mu_1 - \mu_2$.

What is the distribution of $\bar{x}_1$ assuming the sample size is large?

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

What is the distribution of $\bar{x}_2$ assuming the sample size is large?

$$\bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Under the assumption that $\bar{x}_1$ and $\bar{x}_2$ are independent what is the distribution of $\bar{x}_1 - \bar{x}_2$?

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Note: use the formulas for $E(X - Y)$ and $Var(X - Y)$ and the rule that a sum of normal random variables is normal

We can use this information to make inferences about $\mu_1 - \mu_2$ the same way we used knowledge of the distribution of $\bar{x}$ to make inferences about μ.

What sort of things are we going to want to do?

(1) Estimate confidence intervals for $\mu_1 - \mu_2$
(2) Test Hypothesis about $\mu_1 - \mu_2$
(3) Estimate confidence intervals for $p_1 - p_2$
(4) Test Hypothesis about $p_1 - p_2$

Example: From CPS ORG

By race (1987)

```
-------------------------------------------------------------------------------
-> race = 1

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |    4270    710.7766   367.9828        160     1598.4

-------------------------------------------------------------------------------
-> race = 2

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |     471    543.0352   261.4621      187.2     1598.4

-------------------------------------------------------------------------------
-> race = 3

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |     259    556.6517   321.9091        192     1598.4
```

By sex (1987)

```
-------------------------------------------------------------------------------
-> race = 1

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |    4270    710.7766   367.9828        160     1598.4

-------------------------------------------------------------------------------
-> race = 2

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |     471    543.0352   261.4621      187.2     1598.4

-------------------------------------------------------------------------------
-> race = 3

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |     259    556.6517   321.9091        192     1598.4
```

By Married (1987)

```
-------------------------------------------------------------------------------
-> married = 0

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |     902    652.1596   344.1542      187.2     1598.4

-------------------------------------------------------------------------------
-> married = 1

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnwke |    1926    851.3188   385.0304      187.2     1598.4
```

Example: Wisconsin Presidential Polls

Fox News/Opinion Dynamics Polll
October 30-31
700 likely voters, +/-3.5% points
48% 45% 1%
CNN/USA Today/Gallup Poll
October 27-30
1,119 likely voters, +/-3% points
52% 44% 1%
Mason-Dixon Poll
October 26-29
625 likely voters, +/-4% points
46% 48% 1%
Badger Poll
October 23-27
545 likely voters, +/-4.5% points
48% 45% 3%
American Research Group
October 25-27
600 likely voters, +/-4% points
47% 48% 1%
CNN/USA Today/Gallup Poll
October 16-19
678 likely voters, +/- 4% points
50% 44% 3%