2. Review of descriptive and inferential statistics

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Research design for causal inference

- Many (most?) interesting questions in education research and policy are *causal* in nature. Examples from last time: what explains gaps in post-secondary enrollment by race, gender, and income? Do students have better outcomes in charter vs. traditional public schools?
- Estimating causal effects requires a *counterfactual*—what the outcome would have been in a different state of the world. This is unobservable!
- Naive comparisons are fraught with *selection bias*.
- Good research design is about providing convincing counterfactuals.
  - Randomized controlled trials
  - Quasi-experiments

Today

A review of descriptive and inferential statistics:

- Describing data
- Using a sample to make inferences about the population:
  - Sampling distributions
  - Estimation
  - Confidence intervals
  - Hypothesis tests

Descriptive vs. inferential statistics

- Statistical methods can be classified as **descriptive** or **inferential**.
- **Descriptive statistics** are used to *describe* outcomes in a population or sample (e.g., central tendency, variation, distribution shape; overall or by subgroup; correlation).
- **Inferential statistics** are used to make *inferences* or *predictions* about a population larger than that observed in the data.

Descriptive vs. inferential statistics

- **Population**: the universe of outcomes of interest
  - GPAs of all Vanderbilt undergraduates
  - Commuting time for all Vanderbilt graduate students
  - Incomes of U.S. households
  - Math ability of 4th grade students
  - Written language skill of 11-year olds in rural Pakistani villages
- Notice each of these examples includes an *outcome* and a *unit of observation* (and often a time/place)
- The researcher may or may not observe (or be able to observe) the population of interest.

Descriptive vs. inferential statistics

- **Sample**: a subset of the population, chosen at random or by some other method.
- Descriptive statistics can be conducted on either a population or a sample.
- The key difference is how these statistics are used / interpreted.

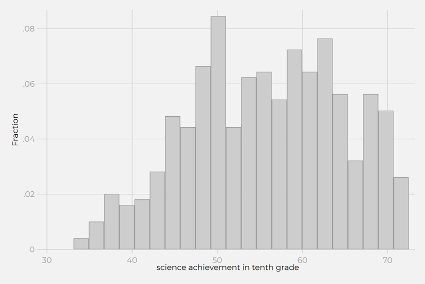
Descriptive vs. inferential statistics

- It may be impossible, or cost-prohibitive, to observe the full population. In these cases a sample can be used to make inferences about the population.
- An important step in inferential statistics is the *quantification of uncertainty* (e.g., standard errors or a "margin of error"). Covered in the second half of this lecture.

Describing data

Describing data

The first step in data analysis is almost always *description*—understanding the *distribution* or *relative frequency* of your variables of interest.



Variable measurement

The tools we use for description depend in part on what *type* of variable it is. Some key characteristics:

- A **quantitative** variable is on a numeric scale, where the numeric values express the magnitude of some property or characteristic.

- A **categorical** variable consists of a number of distinct categories that lack a natural ordering.
  - Special case: a **dichotomous** or **binary** variable has *two* categories (e.g., Y/N, employed or unemployed, graduated or not). Sometimes called a "dummy" or "indicator" variable, and can be coded 0-1.

Variable measurement

Quantitative variables can be discrete or continuous:

- **discrete** or **count**: the variable can take on a *countable* number of values (e.g., values can be represented by integers). Sometimes includes negative numbers, sometimes not.
- **continuous**: the variable can take on a *continuum* of values (e.g., all values between any arbitrary values a and b)

Tonight's sample datasets

We will refer to two datasets tonight (all found on Github):

1. `nels.dta`: an extract from the National Education Longitudinal Study of 1988 (NELS-88). 8th graders followed from 1988 forward.
2. `nyvoucher.dta`: pre and post-test scores from the New York Scholarship Program, a randomized experiment of private school vouchers in NYC (see M&W ch. 4).

Describing categorical variables

Categorical variables lack a natural ordering, so we often describe their distributions—how often different values come up—using a relative frequency table or bar graph.

```
. tabulate parmarl8, missing
```

parents' marital status in eighth grade	Freq.	Percent	Cum.
divorced	26	5.20	5.20
widowed	6	1.20	6.40
separated	8	1.60	8.00
never married	5	1.00	9.00
marriage-like relationship	5	1.00	10.00
married	427	85.40	95.40
.	23	4.60	100.00
Total	500	100.00	

Note: the `missing` option includes missing values as one of the categories. You can leave this off.

Describing categorical variables

Also see the user-created `fre` command which shows the relative frequency with and without missing values:

```
. fre parmar
```

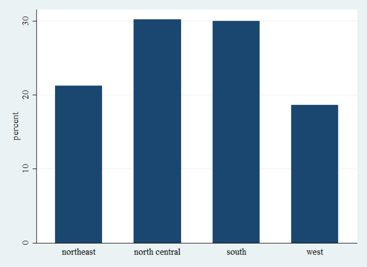
parmarl8 — parents' marital status in eighth grade

		Freq.	Percent	Valid	Cum.
Valid	1 divorced	26	5.20	5.45	5.45
	2 widowed	6	1.20	1.26	6.71
	3 separated	8	1.60	1.68	8.39
	4 never married	5	1.00	1.05	9.43
	5 marriage-like relationship	5	1.00	1.05	10.48
	6 married	427	85.40	89.52	100.00
	Total	477	95.40	100.00	
Missing	.	23	4.60		
Total		500	100.00		

`fre` is also nice in that it shows you both variable *values* and *labels*.

Describing categorical variables

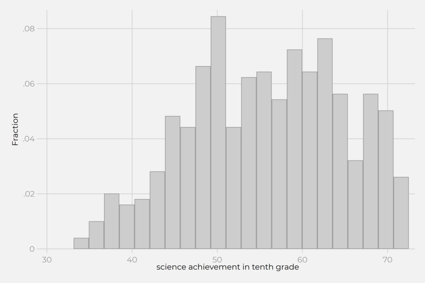
`graph bar (percent), over(`*`region`*`)`



Histogram

Histograms show relative frequency for quantitative variables. They are like bar graphs, where the bars represent *ranges* of observations (bins).

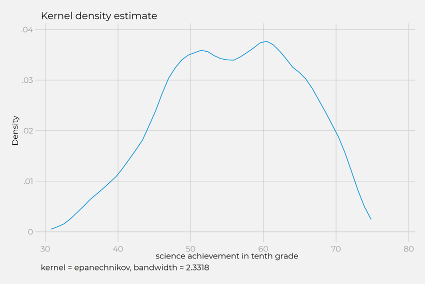
`histogram achsci10, fraction`



Density

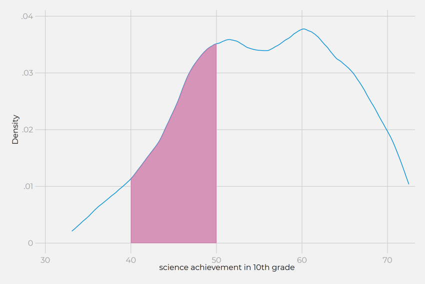
A **density** plot shows you what would happen if the histogram bins get narrower and narrower.

```
kdensity achsci10
```



Density

The area under a density is 1. The area between certain values is the *probability* of being within that range.



Note: about 23% of NELS students scored between 40 and 50.

Summary statistics

A histogram or density plot can tell you just about everything about a variable's distribution. However, you will probably want to summarize it in some useful ways:

- Measures of **central tendency** or **location**: mean, median, percentiles
- Measures of **variability**: range, variance, standard deviation, inter-quartile range (IQR)
- Measures of **skewness**: skewness statistic

Mean

The **mean** adds all of the observed values and divides by the number of observations n.

Let $x_1, x_2, x_3, ..., x_n$ represent the n values of a variable x (x_i is the ith observation, and i is the *index*). The **mean** is:

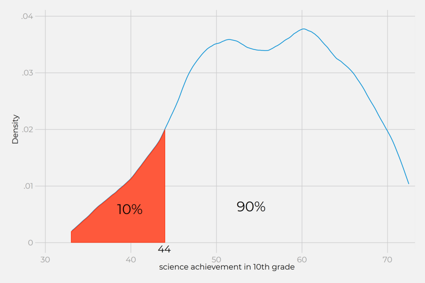
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```
. summ achsci10
```

Variable	Obs	Mean	Std. Dev.	Min	Max
achsci10	497	55.79571	8.96852	33.13	72.48

Percentiles

The *p*th **percentile** of a distribution is the value for which *p*% of observations are less.



A score of 44 is the 10th percentile of the 10th grade science achievement distribution in the NELS. About 10% of students scored below 44.

Percentiles

Common percentiles are reported in the `summarize` command, including the **median** (the 50th percentile).

```
. summ achsci10,detail

                 science achievement in tenth grade
-------------------------------------------------------------
      Percentiles      Smallest
 1%        35.91          33.13
 5%        40.08           34.7
10%        44.07          35.12       Obs                 497
25%        49.02          35.36       Sum of Wgt.         497

50%        56.06                      Mean           55.79571
                        Largest       Std. Dev.       8.96852
75%        62.76          71.72
90%        67.93          71.72       Variance       80.43435
95%        69.51          71.72       Skewness      -.1875573
99%        71.72          72.48       Kurtosis       2.225198
```

Think of the mean as a "representative value" and the median as a "representative observation."

Variance and standard deviation

Measures of variability tell us how "spread out" the data are. The **variance** is the mean of the variable's squared variation around its mean:

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

The variance by itself is hard to interpret, so we take the square root to return to the scale of the original variable. This is the **standard deviation**:

$$s = \sqrt{s^2}$$

z-scores

The **z-score** is another useful measure of location in a variable's distribution:

$$z = \frac{x - \bar{x}}{s}$$

z tells you how many standard deviations away from $\bar{x}$ that a specific value x is. It can tell you "how unusual" x is relative to the amount of variation we typically see in that variable.

We often convert test scores to *z-scores* so that they are on a common scale. It has a mean of 0 and a standard deviation of 1.

Inter-quartile range

Another useful measure of variability is the **inter-quartile range**: the 75th percentile minus the 25th percentile.

```
. summ achsci10,detail

              science achievement in tenth grade
-------------------------------------------------------------
      Percentiles      Smallest
 1%        35.91          33.13
 5%        40.08           34.7
10%        44.07          35.12       Obs                 497
25%        49.02          35.36       Sum of Wgt.         497

50%        56.06                      Mean           55.79571
                        Largest       Std. Dev.       8.96852
75%        62.76          71.72
90%        67.93          71.72       Variance       80.43435
95%        69.51          71.72       Skewness      -.1875573
99%        71.72          72.48       Kurtosis       2.225198
```

IQR = 62.76 - 49.02 = 13.74

Inferential statistics

Inferential statistics

With descriptive statistics, all of the outcomes are known—we are just finding useful ways of summarizing them.

When using statistics for *inferential* purposes, one has to think about a larger population distribution that we don't observe.

- What population distribution "generated" the real-world data we observe? Sometimes called the **data generating process**.
- What is this population distribution's shape (e.g., is it *normal*)? What is its *mean*? Its *median*? Its *variance*?

We use our sample—usually some kind of random sample—to make inferences about the population distribution.

A note about notation

- English/Latin letters represent data (e.g., x or x_i).
- Modifications of these usually represent a calculation using data (e.g., $\bar{x}$ is the sample mean).
- Greek letters represent unknown "true" parameters (e.g., μ is the population mean, σ^2 is the population variance, and σ is the population standard deviation).
- Modification of Greek letters usually represent an estimator of a true parameter (e.g., $\hat{\mu}$).

We typically use $\bar{x}$ as an **estimator** of μ. It is a calculation that (we hope) gives us a good estimate of the "truth."

Quantifying uncertainty

When making statistical inferences, it is important to *quantify your uncertainty*. For example, suppose you calculate the sample average science test score for a sample of students:

$\bar{x} = 55$ with a margin of error of ± 4 points.

The margin of error is useful here because it gives us a sense of "how close" we are likely to be to the true population mean science score (μ).

It also helps us rule out other theoretical distributions. For example, it is unlikely these data came from a distribution with a true population mean score of 75!

Sampling distributions

One of the most important things you learn in statistics:

- Statistics calculated from random samples of the population (such as the sample mean $\bar{x}$) **are also random variables**!
- That is, they take on different values from one sample to the next—called **sampling variation**.
- The (theoretical) distribution of a sample statistic (like $\bar{x}$) is called a **sampling distribution**.

Try this simulation:
`https://istats.shinyapps.io/sampdist_cont/`

Sampling distributions

What this simulation is doing:

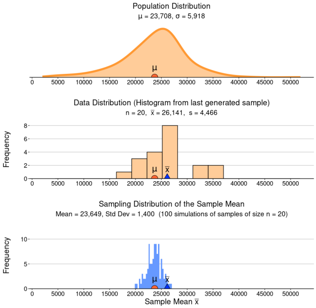
1. Start with a population: median debt—by college—after graduation. ($N = 1,926$)
2. Draw repeated random samples of size n and calculate $\bar{x}$. You choose the sample size and the number of repeated samples.
3. Plot a histogram of the resulting sample means.

Pay attention to the mean and standard deviation of the resulting $\bar{x}$'s!

Note: this is just for illustration! If you really had all of the population data, you would just use it, not sample from it.

Sampling distributions

Here: 100 random samples of size $n = 20$



Sampling distributions

Some key takeaways:

1. On average—across repeated samples—$\bar{x}$ will equal the true population mean. $\bar{x}$ is **unbiased**.
2. Across repeated samples, there is less variation in $\bar{x}$ than there is in the original data. This is more true the larger is n. This is called the **Law of Large Numbers**.
3. The distribution of sample means looks approximately like a **normal distribution**. This is true even if the original data are not normal, if the sample size is large enough. This is the **central limit theorem** at work.

Standard error

The **standard error** is the standard deviation of the sampling distribution. It is a way to "quantify our uncertainty." For $\bar{x}$ this is:

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

In practice we don't know σ, so we estimate it with the sample standard deviation s:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

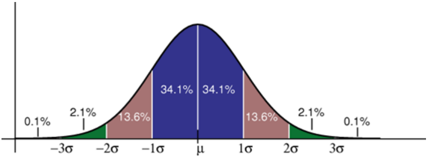
Exercise: NY voucher data

Open the NYSP data and do the following for the pre-test score (*pre_ach*):

1. For sake of this example, keep only the students who did not receive a voucher: (`keep if voucher==0`)
2. Visualize the distribution using a histogram and density plot. What is the shape of the distribution?
3. Calculate the mean, variance, and standard deviation.
4. Recognizing that this sample was one of many samples of $n = 230$ that could have been drawn from this population, calculate the standard error and interpret.

Sampling distributions

The above takeaways about sampling distributions are very powerful, since normal distributions tell us a lot:



- 68.2% of the time, a normal variable will fall within $\pm 1\sigma$ of the mean.
- 95.4% of the time, a normal variable will fall within $\pm 2\sigma$ of the mean.
- 99.6% of the time, a normal variable will fall within $\pm 3\sigma$ of the mean.

Sampling distributions

From the above takeaways, we know that—over repeated samples:

- $\bar{x}$ has an (approximate) normal distribution
- On average, $\bar{x}$ will equal μ
- The standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$ (use s instead of σ)

What this means is that (for example) 68.2% of the time, $\bar{x}$ will fall within ± 1 standard error of μ. 95.4% of the time, $\bar{x}$ will fall within ± 2 SE of μ, and so on. What can we do with this information?

- Confidence intervals
- Hypothesis tests

Confidence intervals

A **confidence interval** for μ is a range of values that will—*in repeated samples*—contain the true population mean some percentage of the time.

- $(1-\alpha)$% is the **confidence level**, the percentage of times in repeated samples that the interval will contain μ:
  - 95% confidence level ($\alpha = 0.05$)
  - 99% confidence level ($\alpha = 0.01$)
  - 90% confidence level ($\alpha = 0.10$)
- α is the **error probability**. We typically choose this.

Common confidence intervals

$$90\% \text{ confidence interval for } \mu : \bar{x} \pm 1.64 * SE(\bar{x})$$

$$95\% \text{ confidence interval for } \mu : \bar{x} \pm 1.96 * SE(\bar{x})$$

$$99\% \text{ confidence interval for } \mu : \bar{x} \pm 2.58 * SE(\bar{x})$$

Exercise: NY voucher data

Continuing with the NYSP data (non-voucher students):

1. Construct 90, 95, and 99% confidence intervals for the true population mean *pre_ach*, and interpret. Do this manually, and using the `mean` estimation command.

Confidence intervals vs. hypothesis tests

Confidence intervals and hypothesis tests are two different approaches to inference.

- Confidence intervals:
  - We have no *a priori* idea of what μ is.
  - The confidence interval represents a range of "likely values" for μ.
  - The width of the interval reflects sampling variability (the standard error).
- Hypothesis tests:
  - Begin with a hypothesis about μ.
  - Ask whether the sample statistic $\bar{x}$ is consistent with this hypothesis.
  - Use knowledge of the *sampling distribution* of $\bar{x}$ to assess how plausible the hypothesis is.
  - Also known as *significance testing*.

Statistical hypotheses

- The **null hypothesis** ($\mathbf{H_0}$): a statement that a population parameter takes a particular value. The null hypothesis is often the *lack* of a hypothesized finding: e.g., "no effect" or "no difference."
- The **alternative hypothesis** ($\mathbf{H_1}$ or $\mathbf{H_a}$): a statement that the population parameter falls in some alternative range of values (contrary to H_0). The alternative hypothesis is often an "effect" or "difference" that the researcher is testing for.

Alternatives can be **one-sided** (directional) or **two-sided** (non-directional).

Example: one-sided hypothesis test

The distribution of math SAT scores in the population has a mean of 500 and $\sigma = 100$. You believe that the mean math SAT score in California is *higher* than the national average (assume σ is the same). To test this hypothesis, you randomly sample n=1,600 California math SAT scores. (Assume we don't have access to the population of California scores).

- Let μ_c represent the population mean SAT in California.
- H_0: $\mu_c = 500$
- H_1: $\mu_c > 500$

The null hypothesis is that $\mu_c = \mu_0 = 500$, i.e., the CA mean is the same as the national average.

Example: one-sided hypothesis test

We compute the sample mean ($\bar{x}$) from our random sample of California scores. If the sample mean is "sufficiently higher" than 500 we *reject* H_0 in favor of H_1. The question is:

- What value of $\bar{x}$ is "sufficiently higher?"

- What value of $\bar{x}$ would make H_0 seem *implausible*?

Example: one-sided hypothesis test

In hypothesis testing, we begin with the assumption H_0 is true. In this example, *if* H_0 *were true*, sample means calculated from random samples of size $n = 1,600$ will:

- follow a normal distribution
- have a mean of 500
- have a standard error of $\sigma/\sqrt{n} = 100/\sqrt{1600}$, or 2.5

Note: we are assuming σ is known. If we didn't, we could use s.

Example: one-sided hypothesis test

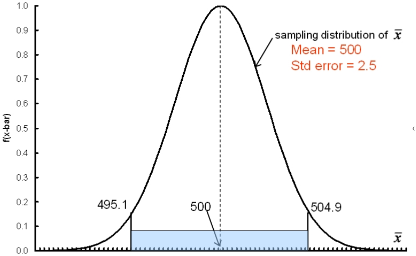


Figure: Sampling distribution of $\bar{x}$ under H_0

Example: one-sided hypothesis test

If H_0 were true, then 95% of the time $\bar{x}$ calculated from a random sample of $n = 1,600$ will fall between:

$$\mu_0 \pm 1.96(\sigma/\sqrt{n})$$

$$500 \pm 1.96(2.5) = (495.1, 504.9)$$

Any realized $\bar{x}$ in this interval wouldn't be that unusual. What about a realized sample mean of $\bar{x} = 507$?

Note: the interval above looks like a confidence interval, but it is centered at μ_0, not $\bar{x}$.

Example: one-sided hypothesis test

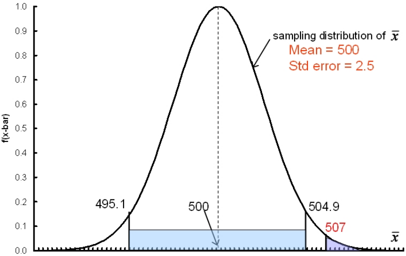


Figure: Sampling distribution of $\bar{x}$ under H_0, and a realized sample mean of 507

Example: one-sided hypothesis test

507 is $(507 - 500)/2.5 = 2.8$ standard errors above μ_0. The probability of obtaining a sample mean of 507 or higher, assuming H_0 is true, is:

$$Pr(z > 2.8) = 0.0026$$

In other words, *very unlikely*. The value 2.8 is called a **test statistic**. 0.0026 is called the ***p*-value**.

The *p*-value comes from the normal distribution. In Stata: `display 1-normal(2.8)`

Example: one-sided hypothesis test

What about a realized sample mean of $\bar{x} = 502$? 502 is $(502 - 500)/2.5 = 0.8$ standard errors above μ_0. The probability of obtaining a sample mean of 502 or higher, assuming H_0 is true, is:

$$Pr(z > 0.8) = 0.2119$$

In other words, not that unlikely. The value 0.8 is our test statistic, and 0.2119 is the *p*-value.

p-values

In general a ***p*-value** is the probability that a test statistic equals the realized value *or a value even more extreme* in the direction predicted by H_1 if H_0 is true.

- Above, the probability of obtaining an $\bar{x}$ of 507 or higher was 0.0026.
- Above, the probability of obtaining an $\bar{x}$ of 502 or higher was 0.2119.
- Intuitively, if we obtain an $\bar{x}$ that would be highly unlikely if H_0 were true (such as 0.0026) then H_0 *was probably not true to begin with.*
- The *less consistent* is the test statistic with H_0, the *smaller* is the p-value.
- A small enough p-value will lead us to reject H_0.

Significance levels

The researcher decides ahead of time the threshold p-value at which she will conclude the evidence is sufficiently strong against H_0.

- The threshold value is called the **significance level** of the test, or α.
- Some common significance levels are 0.05, 0.01, 0.10
- When $p < 0.05$ we say the result is "significant at the 0.05 level."
- When $p < 0.01$ we say the result is "significant at the 0.01 level," etc.
- The *higher* is α, the "lower the bar" for rejection of H_0.
- The *smaller* is α, the "higher the bar" for rejection.

Exercise: NY voucher data

Hypothesis tests are easily conducted in Stata using `ttest`. Continuing with the NYSP data (non-voucher students):

1. Test the one-sided null hypothesis $H_0 : \mu = 23$
2. Test the one-sided null hypothesis $H_0 : \mu = 21$
3. Test the two-sided null hypothesis $H_0 : \mu = 20$

In each case, report the p-value and interpret. Use $\alpha = 0.05$. Note, for two-sided tests the p-value is the p for the one-sided tests $\times 2$.

More on confidence intervals vs. hypothesis tests

Confidence intervals can be used to conduct 2-sided hypothesis tests! A 95% confidence interval (for example) gives you the set of null hypotheses that would <u>not</u> be rejected at the 0.05 level.

Statistical tests for comparing two groups

The most interesting hypothesis tests are those comparing two (or more) groups. For example:

- Do female executives earn less on average than males?
- Do 4th graders in an experimental reading program perform differently on standardized reading tests than 4th graders not in the program?
- Are women more likely to vote for Democratic candidates than men?
- Do subjects participating in a 6-week weight loss program lose more weight over time than those who do not participate in the program?
- Has obesity among children aged 10-12 increased between 2010 and 2020?
- Were COVID infection rates higher in counties without a mask mandate than counties with them?

Statistical tests for comparing two groups

The steps for conducting a test comparing two means are the same as those for the test of a single mean. The most common null hypothesis is that there is *no difference* between the two population means:

$$H_0 : \mu_1 = \mu_2$$

Equivalently,

$$H_0 : \mu_2 - \mu_1 = 0$$

The alternative H_1 is that there *is* a difference.

Note: it doesn't matter which mean you subtract from the other, as long as you keep them straight.

Hypothesis test steps

Hypothesis tests proceed like this:

1. Determine H_0 and H_1.
2. Determine what test statistic you will use.
3. Determine the sampling distribution of your test statistic under H_0. (This requires knowing the standard error.
4. Determine the probability of obtaining your *observed* test statistic if H_0 is true (the p-value), and draw a conclusion.

Statistical tests for comparing two groups

To test H_0 above, you will use the *sample* difference in means: $\bar{x}_2 - \bar{x}_1$. Because these come from random samples, this difference in means is also random, and will vary from sample to sample. What we know:

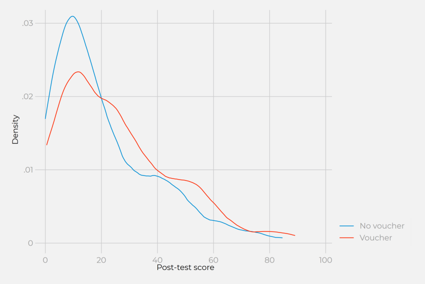
1. The sampling distribution of $\bar{x}_2 - \bar{x}_1$ has a mean of $\mu_2 - \mu_1$
2. The sampling distribution is approximately normal (with large enough sample sizes)
3. The standard error of $\bar{x}_2 - \bar{x}_1$ is:

$$SE(\bar{x}_2 - \bar{x}_1) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

As before, we can substitute s for σ.

Exercise: NY voucher data

Did the voucher "work?" Let's compare the mean post-test scores in the NYSP data:



Exercise: NY voucher data

Now formally test the hypothesis that the population means are equal:

$$H_0 : \mu_V - \mu_{NV} = 0$$

$$H_1 : \mu_V - \mu_{NV} \neq 0$$

Obtain the p-value and interpret. Note Stata will also provide a confidence interval for the difference in means.

Problem set 1

Problem set 1 will give you practice with many of these concepts:

- Descriptive statistics
- Confidence intervals for one mean
- Hypothesis tests for a difference in means

Next time

Multiple regression fundamentals

- Read: Stock & Watson chapter 4