MATH 1052H - S62 - Lecture 01

Review

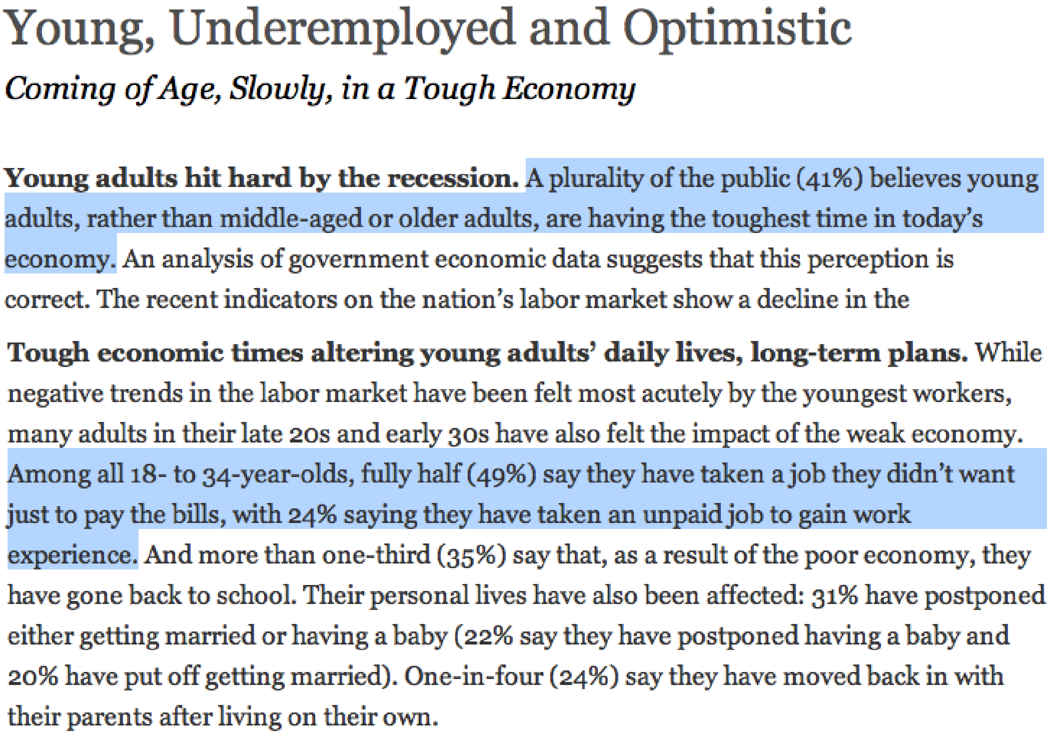
Material Reviewed

In the first two lectures, we will review key ideas from MATH 1051H that you are expected to know: standard error, hypothesis tests, and confidence intervals. Mentioned, but not explicitly reviewed, will also be earlier material on distributions and probabilities.

Variability in estimates

The idea of Standard Error

In this brief unit, we will review the idea of sampling variability. The key concept is that if you have a sample (data!), then computing **statistics** on that data results in a random variables. And random variables have distributions.



Young, Underemployed and Optimistic

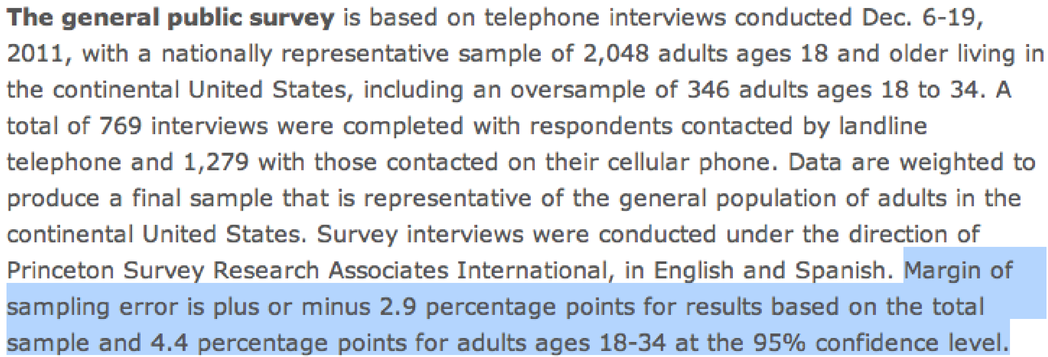
Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

http://pewresearch.org/pubs/2191/young-adults-workers-labor-market-pay-careers-advancement-recession

Margin of error



The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today’s economy.
- 49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn’t want just to pay the bills.

Parameter estimation

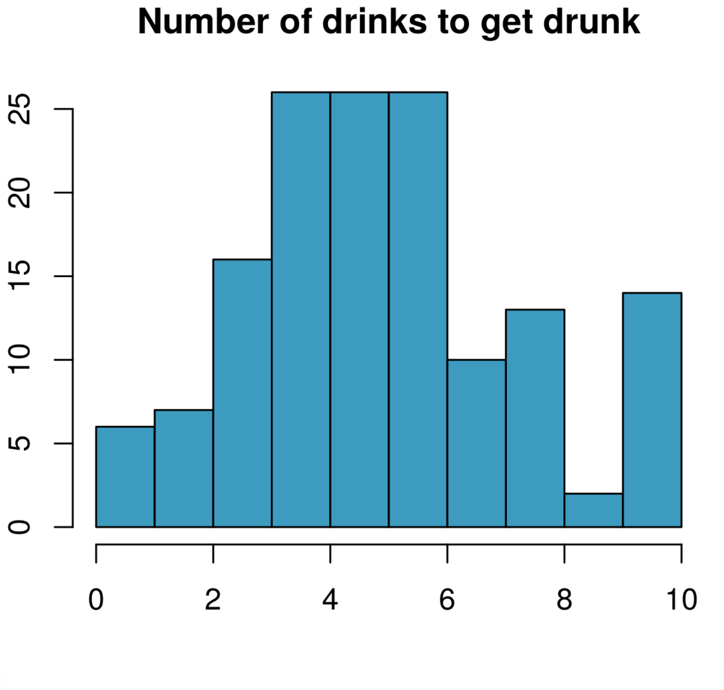
- We are often interested in **population parameters**.
- Since complete populations are difficult (or impossible) to collect data on, we use **sample statistics** as **point estimates** for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the **margin of error** associated with our point estimate.
- But before we get to quantifying the variability among samples, let’s try to understand how and why point estimates vary from sample to sample.

Parameter estimation

Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

The following histogram shows the distribution of number of drinks it takes a group of college students to get drunk. We will assume that this is our population of interest. If we randomly select observations from this data set, which values are most likely to be selected (which are least likely)?



Suppose that you don’t have access to the population data. In order to estimate the average number of drinks it takes these college students to get drunk, you might sample from the population and use your sample mean as the best guess for the unknown population mean.

- Sample, with replacement, ten students from the population, and record the number of drinks it takes them to get drunk.
- Find the sample mean.
- Plot the distribution of the sample averages obtained by members of the class.

1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Example: List of random numbers: 59, 121, 88, 46, 58, 72, 82, 81, 5, 10

1	7	16	3	31	5	46	4	61	10	76	6	91	4	106	6	121	6	136	6
2	5	17	10	32	9	47	3	62	7	77	6	92	0.5	107	2	122	5	137	7
3	4	18	8	33	7	48	3	63	4	78	5	93	3	108	5	123	3	138	3
4	4	19	5	34	5	49	6	64	5	79	4	94	3	109	1	124	2	139	10
5	6	20	10	35	5	50	8	65	6	80	5	95	5	110	5	125	2	140	4
6	2	21	6	36	7	51	8	66	6	81	6	96	6	111	5	126	5	141	4
7	3	22	2	37	4	52	8	67	6	82	5	97	4	112	4	127	10	142	6
8	5	23	6	38	0	53	2	68	7	83	6	98	4	113	4	128	4	143	6
9	5	24	7	39	4	54	4	69	7	84	8	99	2	114	9	129	1	144	4
10	6	25	3	40	3	55	8	70	5	85	4	100	5	115	4	130	4	145	5
11	1	26	6	41	6	56	3	71	10	86	10	101	4	116	3	131	10	146	5
12	10	27	5	42	10	57	5	72	3	87	5	102	7	117	3	132	8		
13	4	28	8	43	3	58	5	73	5.5	88	10	103	6	118	4	133	10		
14	4	29	0	44	6	59	8	74	7	89	8	104	8	119	4	134	6		
15	6	30	8	45	10	60	4	75	10	90	5	105	3	120	8	135	6		

Sample mean: $\frac{8+6+10+4+5+3+5+6+6+6}{10} = 5.9$

Sampling distribution

What you just constructed is called a *sampling distribution*.

What is the shape and center of this distribution? Based on this distribution, what do you think is the true population average?

Sampling distribution

What you just constructed is called a *sampling distribution*.

What is the shape and center of this distribution? Based on this distribution, what do you think is the true population average?

Approximately 5.39, the true population mean.

Sampling distributions - via CLT

Central limit theorem

Central limit theorem The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim \mathcal{N}\left(\text{mean} = \mu, \text{SE} = \frac{\sigma}{\sqrt{n}}\right),$$

where SE is represents **standard error**, which is defined as the standard deviation of the sampling distribution. If σ is unknown, use s (recall: standard deviation of sample).

Central limit theorem

- It wasn’t a coincidence that the sampling distribution we saw earlier was symmetric, and centered at the true population mean.
- We won’t go through a detailed proof of why $SE = \frac{\sigma}{\sqrt{n}}$, but note that as n increases SE decreases.
    - As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

CLT - conditions

Certain conditions must be met for the CLT to apply:

- **Independence:** Sampled observations must be independent. This is difficult to verify, but is more likely if
    - random sampling/assignment is used, and
    - if sampling without replacement, $n <$ 10% of the population.

CLT - conditions

Certain conditions must be met for the CLT to apply:

- **Independence:** Sampled observations must be independent. This is difficult to verify, but is more likely if
    - random sampling/assignment is used, and
    - if sampling without replacement, $n <$ 10% of the population.
- **Sample size/skew:** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large. This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.
    - the more skewed the population distribution, the larger sample size we need for the CLT to apply
    - for moderately skewed distributions $n > 30$ is a widely used rule of thumb

Summary

So from this set of ideas, we get the concept that when we have a random sample, we have something called the **standard error**, which is the variation of the sampling distribution under the CLT.

This is a very, very important idea for the following topic.

Hypothesis Testing

Hypothesis Tests as a Trial

Hypothesis testing is very much like a court trial.

- H_0: defendent is innocent (English common law; Justinian Codes, UN Declaration of Human Rights), *versus*
  H_A: defendent is guilty
- We then present the evidence — collect data
- Then we judge the evidence: “Could these data plausibly have happened by chance if the null hypothesis were true?”
    - If they were very unlikely to have occurred, then the evidence raises more than *a reasonable doubt* in our minds about the null hypothesis
- Ultimately, we must make a decision: how unlikely is **unlikely**?

Image from http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zzaqbboq25oevvbg99wpot

A Hypothesis Test as a Trial (continued)

- If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of “not guilty”.
  - The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  - The defendant may, in fact, be innocent, but the jury has no way of being sure.
- Said statistically, we **fail to reject the null hypothesis**.
  - We never declare the null hypothesis to be true, because we simply do not know whether it’s true or not.
  - Therefore we never “accept the null hypothesis”.

A Hypothesis Test as a Trial (continued)

- In a trial, the burden of proof is on the prosecution.
- In a hypothesis test, the burden of proof is on the unusual claim.
- The null hypothesis is the ordinary state of affairs (the status quo), so it’s the alternative hypothesis that we consider unusual and for which we must gather evidence.

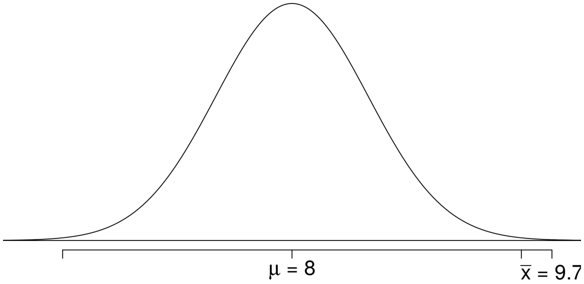
Recap: Concept of Hypothesis Testing

- We start with a **null hypothesis** (H_0) that represents the status quo.
- We also have an **alternative hypothesis** (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation (today) or theoretical methods (later in the course).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Formal testing using p-values

Test Statistic $\bar{x}$, Large Samples

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the test statistic.



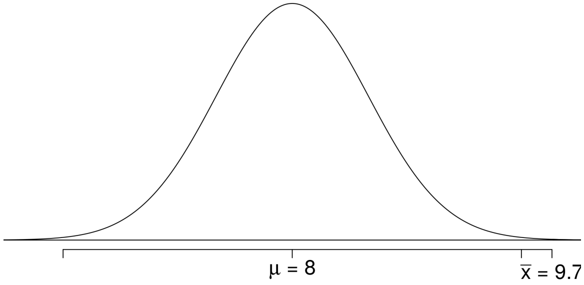
$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} \approx 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result **statistically significant**?

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the test statistic.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} \approx 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result **statistically significant**?

Yes, and we can quantify how unusual it is using a p-value.

p-values

- We then use this test statistic to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

p-values

- We then use this test statistic to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is **low** (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence **reject** H_0.

p-values

- We then use this test statistic to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is **low** (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence **reject** H_0.
- If the p-value is **high** (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence **do not reject** H_0.

A poll by the National Sleep Foundation (USA) found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep **less than** 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

- Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.
- Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.
- Reject H_0, the data prove that college students sleep more than 7 hours on average.
- Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.
- Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Two-sided hypothesis testing with p-values

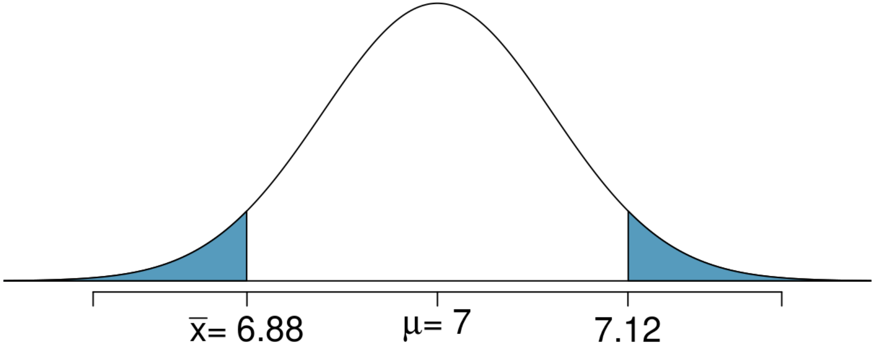
- If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is **different** than the national average?”, the alternative hypothesis would be different.

$$H_0 : \mu = 7$$
$$H_A : \mu \neq 7$$

Two-sided hypothesis testing with p-values

- If the research question was "Do the data provide convincing evidence that the average amount of sleep college students get per night is **different** than the national average?", the alternative hypothesis would be different.
- Then the p-value **would change as well**:

$$\text{p-value} = 0.0485 \times 2 = 0.097$$



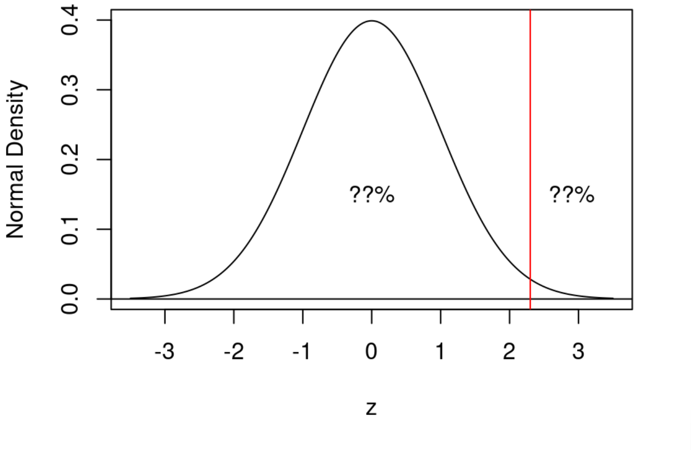
Computing the *p*-value

How do we actually compute the *p*-value? We use **pnorm()**! There's a reason we made you learn about it!

Example: the **test statistic** is 2.3, with hypotheses

$$H_0 : \mu = 5 \qquad \text{versus} \qquad H_A : \mu > 5$$

What is the *p*-value?



Example, continued

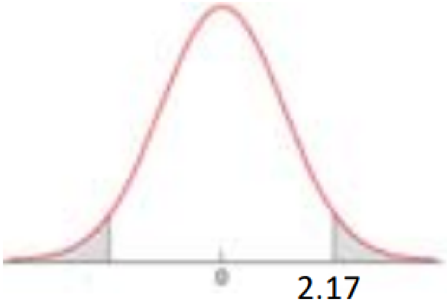
```r
pnorm(2.3, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 0.01072411
```

So the *p*-value for this **one-tailed hypothesis test** is 0.011. What does this imply?

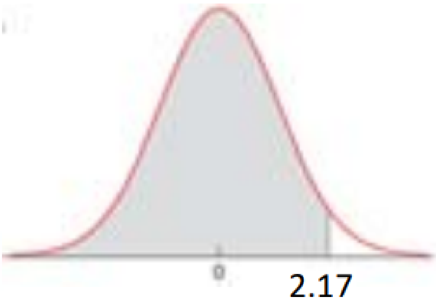
Since $0.011 < 0.05$, we do have evidence at the 95% level to reject the null hypothesis (whatever it is in context), and conclude that $\mu > 5$.

The Alternative Hypothesis …



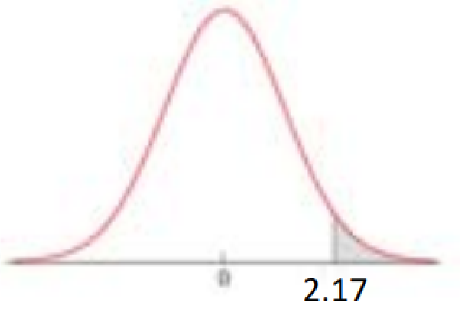
Two-tailed H_A
($\neq$)

Find the area to the right of z and multiply by 2, (or to the left of z if *z* were negative and multiply by 2).



Left-tailed H_A
($<$)

Find the area to the left of z.



Right-tailed H_A
($>$)

Find the area to the right of z.

Inference for other estimators

Inference for other estimators

- The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample **proportions** is also nearly normal when n is sufficiently large (more on this later)

Inference for other estimators

- The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample **proportions** is also nearly normal when n is sufficiently large (more on this later)
- An important assumption about point estimates is that they are **unbiased**, i.e., the sampling distribution of the estimate is centered at the true population parameter it estimates.

Inference for other estimators

- The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample **proportions** is also nearly normal when n is sufficiently large (more on this later)
- An important assumption about point estimates is that they are **unbiased**, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimates.
    - An unbiased estimate does not naturally over or underestimate the parameter. Rather, it tends to provide a “good” estimate.
    - The sample mean is an example of an unbiased point estimate, as are each of the examples we introduce in this section.

Inference for other estimators

- The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample **proportions** is also nearly normal when n is sufficiently large (we'll talk about this in detail in two weeks).
- An important assumption about point estimates is that they are **unbiased**, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimates.
    - That is, an unbiased estimate does not naturally over or underestimate the parameter. Rather, it tends to provide a “good” estimate.
    - The sample mean is an example of an unbiased point estimate, as are each of the examples we introduce in this section.
- Some point estimates follow distributions other than the normal distribution, and some scenarios require statistical techniques that we haven't covered yet - we will discuss most of these in the next course (1052H)

Non-normal point estimates

- We may apply the ideas of hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:
    - the sample size is too small for the normal approximation to be valid;
    - the standard error estimate may be poor; or
    - the point estimate tends towards some distribution that is not the normal distribution.

Non-normal point estimates

- We may apply the ideas of hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:
  - the sample size is too small for the normal approximation to be valid;
  - the standard error estimate may be poor; or
  - the point estimate tends towards some distribution that is not the normal distribution.
- For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for hypothesis testing to these alternative distributions.

When to retreat

- Statistical tools rely on the following two main conditions:
  - **Independence**: A random sample from less than 10% of the population ensures independence of observations. In experiments, this is ensured by random assignment. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.
  - **Sample size and skew**: For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail.
- Whenever conditions are not satisfied for a statistical technique:
  - Learn new methods that are appropriate for the data.
  - **Consult a statistician.**
  - **Ignore the failure of conditions.** This last option effectively invalidates any analysis and may discredit novel and interesting findings.

The *t* Distribution

A Specific Example of Non-Normal

In the last discussion, we talked about how we might approach certain problems where the **normal assumption** does not hold. We're now going to start looking at a specific, famous example of this kind of problem - the *t* distribution.

Example

Mercury in seafood due to pollution is a known problem, especially in heavy industrial areas, although mercury has spread a long way from explicit polluters. Japan as a country consumes a large amount of seafood, and researches were interested in the average mercury content in Rossi's dolphins from the Taiji area. They analyzed 19 dolphins' muscles for mercury content.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Measurements are in micrograms of mercury per wet gram of muscle (μg/wet g).

So, a "begged question": could we do a hypothesis test on this data using what we know so far (e.g., a Z distribution)?

Review: Purpose of Large Sample

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that:

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error (SE), as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

The CLT, which states that sampling distributions will be nearly normal, holds true for any sample size as long as the population distribution is nearly normal.

While this is a helpful special case, it's inherently difficult to verify normality in small data sets.

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.

For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

In Context (Dolphins)

How big is our sample? And, given the summary, how symmetric is our data?

- only 19 samples
- no population σ
- data seems mostly symmetric

The t Distribution

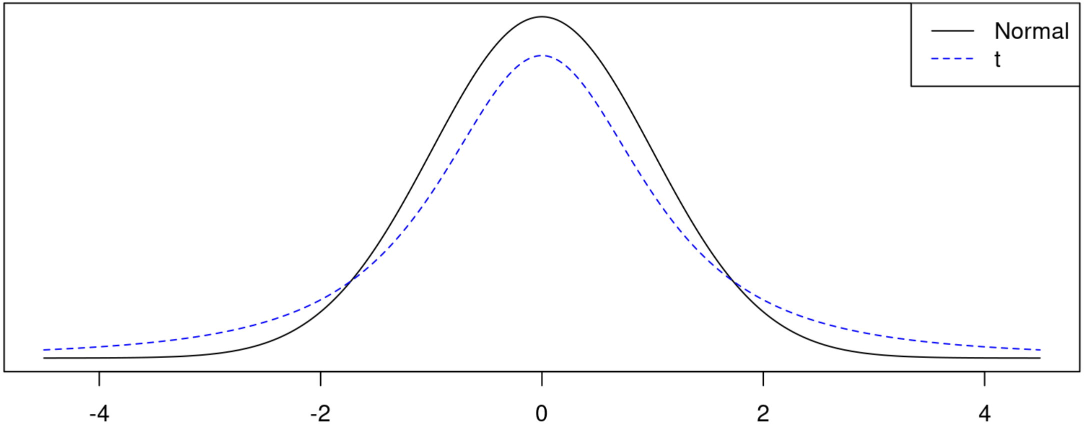
When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.

This distribution also has a bell shape, but its tails are thicker than the normal model's.

Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.

These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)

A plot of t versus $\mathcal{N}$



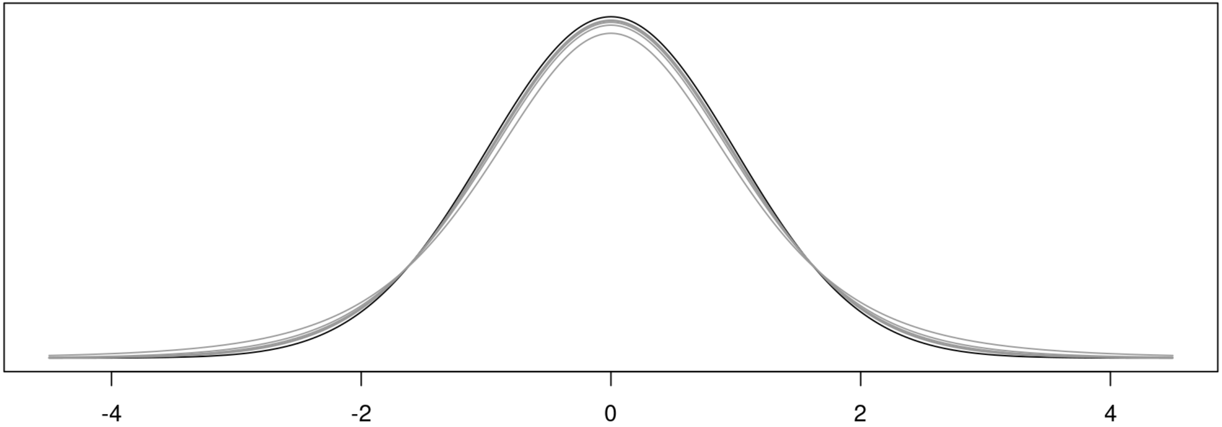
The t Distribution (ctd.)

Always centered at zero, like the standard normal (z) distribution.

Has a single parameter: degrees of freedom (df) – like χ^2.

What happens to the shape of the t distribution as df increases?

The t Distribution (ctd.)



As $df \longrightarrow \infty$, the t distribution approaches the normal!

Asymptotic

What df is required to give arbitrary decimal agreement between the t and z curves? (on a restricted domain)

- 2 decimals: $df = 14$
- 3 decimals: $df = 136$
- 4 decimals: $df = 1370$

What do we usually ask for? 30 df corresponds to 3 decimals for the $[-3, 3]$ domain, which is good enough. So once $df > 30$, people often just use a z instead.

Recap: Inference using a small sample mean

If $n < 30$, sample means follow a t distribution with $\mathrm{SE} = \frac{s}{\sqrt{n}}$, **unless** you are positive you know the population standard deviation σ.

1. **Conditions**:
   - independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
   - $n < 30$ and no extreme skew
2. **Hypothesis Testing**:

$$t_{\mathrm{df}} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1.$$

Back to the Dolphins

Researchers want to know if the average mercury content in these dolphins exceeds 4 μg/wet g. Perform a hypothesis test to answer this question.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Hypothesis Test

$$H_0 : \mu \leq 4 \qquad \text{versus} \qquad H_A : \mu > 4$$

Conditions:

- we assume independence of observations
- $n < 30$
- don’t know σ

So we need to use the t!

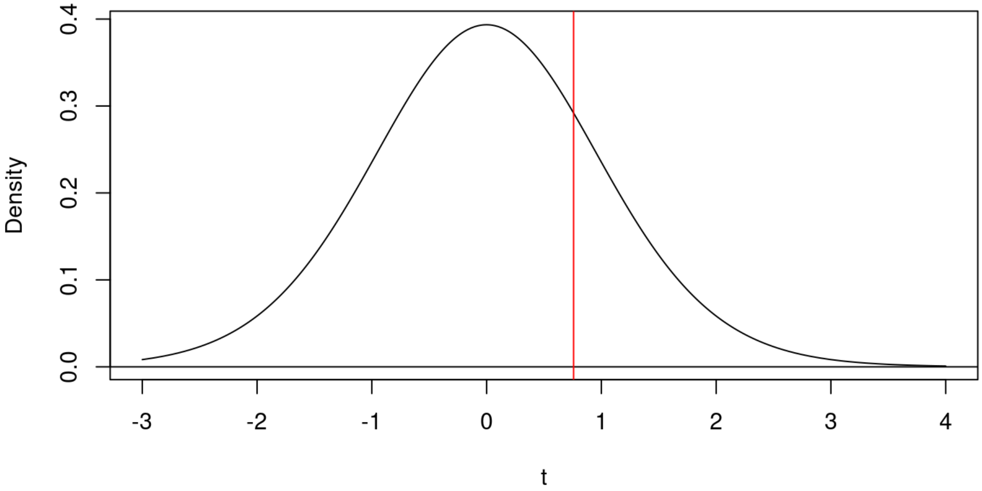
Test Statistic

The test statistic doesn't change: it's a test statistic on the mean, so the **statistic** stays the same! The only difference is that the random variable we get is no longer assumed to be a Z … instead, it's a t.

$$t = \frac{\bar{x} - \mu_0}{\mathrm{SE}_{\bar{x}}} = \frac{4.4 - 4}{\frac{2.3}{\sqrt{19}}} = 0.7581.$$

Computing the p-value

Our alternative is $\mu > 4$, so our p-value goes **up** …



p-value

But there's a trick … when you specify a *t* distribution, you don't specify mean/SD … you have to specify the **degrees-of-freedom** (df). Our rule for the mean is that df is **n-1**: one less than the number of samples you have.

```
pt(q = 0.7581, df = 19 - 1, lower.tail = FALSE)
```

```
## [1] 0.2291013
```

Thus our p-value is 0.229, which is "big", so we would **fail to reject the null** hypothesis, meaning we cannot conclude that the mean is greater than 4 μg/wet g.

Summary of Tests on Means for t Distributions

- Similar assumptions to the Z
- Same test statistic
- Notice the df argument (n-1)
- Same way of computing p-value, except use **pt()** not **pnorm()**
- Same interpretation

Inference for a Single Proportion

Practice

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

1. All 1000 get the drug
2. 500 get the drug, 500 don’t

Practice

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

1. All 1000 get the drug
2. *500 get the drug, 500 don't*

Results from the GSS

The General Social Survey (GSS) collects information and keep a historical record of the concerns, experiences, attitudes, and practices of residents of the United States. Since 1972, the GSS has been monitoring societal change and studying the growing complexity of American society. Canada has been running a similar survey since 1985.

The GSS asks the question from the previous slide. Below is the distribution of responses from the 2010 survey:

All 1000 get the drug	99
500 get the drug, 500 don’t	571
Total	670

http://www.statcan.gc.ca/pub/89f0115x/89f0115x2013001-eng.htm

Parameter and Point Estimation

We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e., would answer “500 get the drug, 500 don’t”? What are the parameter of interest and the point estimate?

Parameter and Point Estimation

We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e., would answer “500 get the drug, 500 don’t”? What are the parameter of interest and the point estimate?

Parameter of Interest: Proportion of **all** Americans who have good intuition about experimental design

p: a population proportion

Parameter and Point Estimation

We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e., would answer “500 get the drug, 500 don’t”? What are the parameter of interest and the point estimate?

Parameter of Interest: Proportion of **all** Americans who have good intuition about experimental design

p: a population proportion

Point Estimate: proportion of **sampled** Americans who have good intuition about experimental design.

$\hat{p}$: a sample proportion

Inference on a Proportion

What percent of all Americans have good intuition about experimental design, i.e., would answer “500 get the drug 500 don’t”?

We can answer this question using a confidence interval, which we know is always of the form

$$\text{point estimate} \pm \text{ME}$$

where

$$\mathrm{ME} = z^\star \times \mathrm{SE}. \text{ (or } t^\star \times \mathrm{SE}\text{)}$$

So what is the SE of our point estimate, $\mathrm{SE}_{\hat{p}}$?

New Formula: SE of a Point Estimate $\hat{p}$

When we have a **sample proportion**, the standard error has a known formula:

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

What are p and n?

1. n is the number of samples (it’s a **sample proportion**)
2. p is the true underlying population proportion …

But we don’t know p!

We “cheat” here, and replace p with $\hat{p}$. It mostly works.

Sample Proportions are Almost Normally Distributed

Remember the Central Limit Theorem (CLT).

Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\mathrm{SE}_{\hat{p}}$ from the last slide. We can write this formally.

$$\hat{p} \sim \mathcal{N}\left(\text{mean} = p, \mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}\right)$$

But, of course, this is only true under certain conditions … *any guesses?*

Sample Proportions are Almost Normally Distributed

Remember the Central Limit Theorem (CLT) from earlier.

Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\mathrm{SE}_{\hat{p}}$ from the last slide. We can write this formally.

$$\hat{p} \sim \mathcal{N}\left(\text{mean} = p, \mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}\right)$$

But, of course, this is only true under certain conditions … *any guesses?*

The requirements of the CLT! Independent observations, and "enough" samples

Rule of Thumb for Proportions

There is a rule of thumb for what “enough samples” means for a sample proportion inference:

1. At least 10 success cases
2. At least 10 failure cases

If you do not have the above, the CLT may not be a good approximation.

Back to the GSS Question

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Back to the GSS Question

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Given:

- $n = 670$
- $\hat{p} = 0.852$

Check the conditions!

Back to the GSS Question

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Given:

- $n = 670$
- $\hat{p} = 0.852$

Check the conditions!

1. **Independence**: The GSS is sampled randomly, and the population is much larger than the sample, so we can assume the responses are random.

Back to the GSS Question

The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Given:

- $n = 670$
- $\hat{p} = 0.852$

Check the conditions!

1. **Independence**: The GSS is sampled randomly, and the population is much larger than the sample, so we can assume the responses are random.
2. **Enough Samples**: 571 people answered correctly (success) and 99 answered incorrectly (failure). Both numbers are greater than 10.

Practice

We are given $n = 670$, $\hat{p} = 0.852$, and we know that

$$\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Which of the following is the correct calculation of the 95% confidence interval?

1. $0.852 \pm 1.96 \times \sqrt{\frac{0.85(0.15)}{670}}$
2. $0.852 \pm 1.65 \times \sqrt{\frac{0.85(0.15)}{670}}$
3. $0.852 \pm 1.96 \times \frac{0.85(0.15)}{\sqrt{670}}$
4. $571 \pm 1.96 \times \frac{571(99)}{670}$

Practice

We are given $n = 670$, $\hat{p} = 0.852$, and we know that

$$\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Which of the following is the correct calculation of the 95% confidence interval?

1. $0.852 \pm 1.96 \times \sqrt{\frac{0.85(0.15)}{670}} = (0.825, 0.879)$
2. $0.852 \pm 1.65 \times \sqrt{\frac{0.85(0.15)}{670}}$
3. $0.852 \pm 1.96 \times \frac{0.85(0.15)}{\sqrt{670}}$
4. $571 \pm 1.96 \times \frac{571(99)}{670}$

Recap: Inference for One Proportion

Population parameter p, Point Estimate $\hat{p}$

Recap: Inference for One Proportion

Population parameter p, Point Estimate $\hat{p}$

Conditions:

- independence
  - random sample, and less than 10% of population
- at least 10 successes and 10 failures
  - if not, we can't use the normal approximation $\rightarrow$ use randomization/permutation instead

Recap: Inference for One Proportion

Population parameter p, Point Estimate $\hat{p}$

Conditions:

- independence
  - random sample, and less than 10% of population
- at least 10 successes and 10 failures
  - if not, we can't use the normal approximation $\rightarrow$ use randomization/permutation instead
- Standard Error (SE) = $\sqrt{\frac{p(1-p)}{n}}$
  - for CI, use $\hat{p}$
  - for HT, use p_0

Example: Libraries

Do the majority of voters in a large city favour increased funding for public libraries? Suppose a poll of 250 randomly selected voters in this city found that 140 of them favoured increased funding for public libraries.

Hypotheses:

$$H_0 : p = 0.50 \qquad \text{versus} \qquad H_A : p > 0.50$$

Check assumptions

- The sample is random and independent.
- Large sample size: $np_0 = n(1 - p_0) = 250(0.5) = 125 \geq 10$
- Large population: the city is large, so the number of voters is at least $10 \times 250 = 2500$

Computing the test statistic

Start with:

$$\hat{p} = \frac{140}{250} = 0.56$$

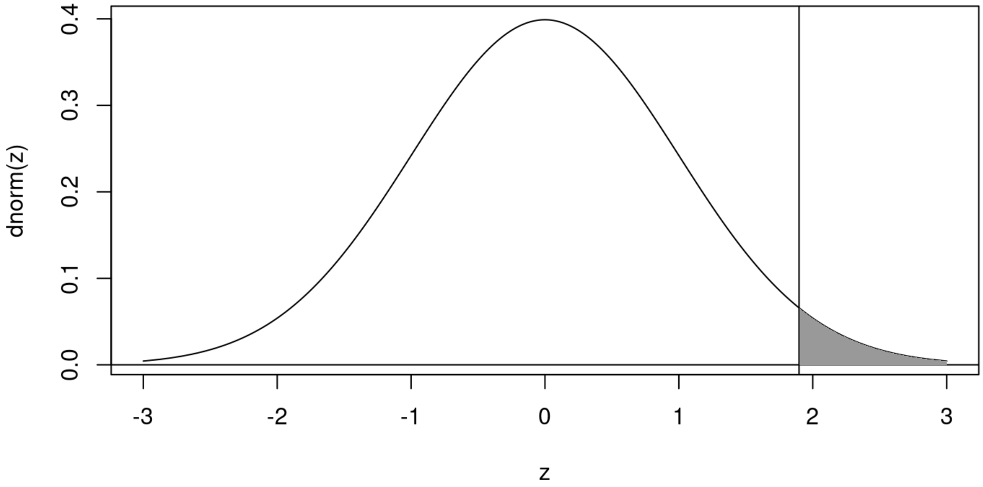
and then compute the SE:

$$\text{SE} = \sqrt{\frac{0.5(0.5)}{250}} = 0.031623.$$

Combine these to make the test statistic:

$$Z_{\text{test}} = \frac{0.56 - 0.50}{0.031623} = 1.8974$$

Visualizing the *p*-value



Computing the *p*-value

```r
1 - pnorm(q = 1.8974)
```

```
## [1] 0.02888758
```

```r
pnorm(q = 1.8974, lower.tail = FALSE)
```

```
## [1] 0.02888758
```

Conclusion

So, since $p < \alpha$ (0.05), we reject the null.

Thus, we reject the null hypothesis, and conclude that there is evidence to support a majority (more than half) of voters in a large city favouring increased funding for libraries.