Chapter 4: Foundations for inference

Chapter 4

Foundations for inference

Some slides developed by Mine Çetinkaya-Rundel of OpenIntro

Variability in estimates

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

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.

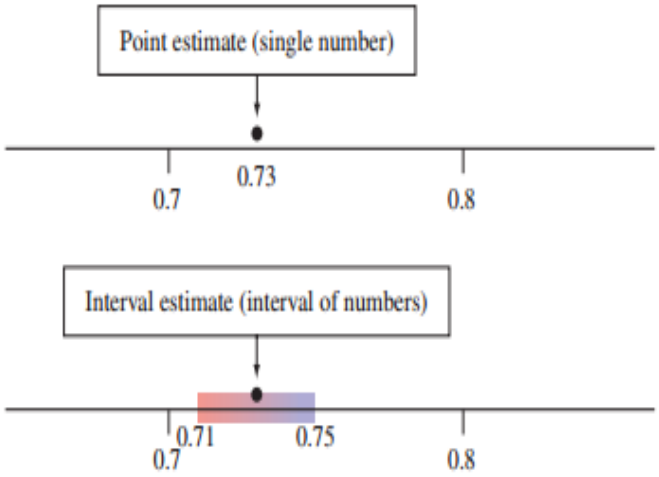
Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Not the same, but only somewhat different.

Point Estimate and Interval Estimate

- A point estimate is a single number that is our “best guess” for the parameter
- An interval estimate is an interval of numbers within which the parameter value is believed to fall.

- 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.



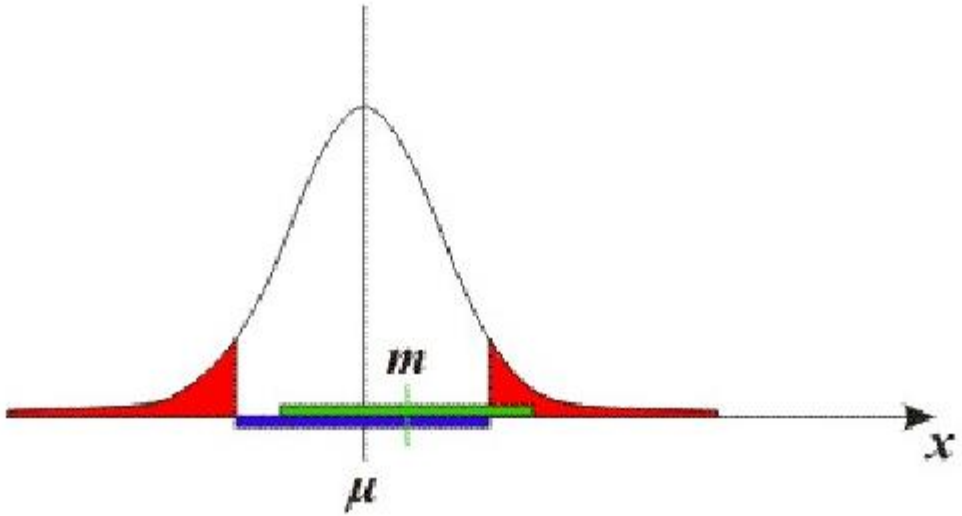
A Point Estimate Predicts a Parameter by a Single Number. An interval estimates an interval of numbers that are believable values for the parameter

Point Estimate and sample statistics (Interval Estimate)

A point estimate doesn't tell us how close the estimate is likely to be to the parameter

An interval estimate is more useful

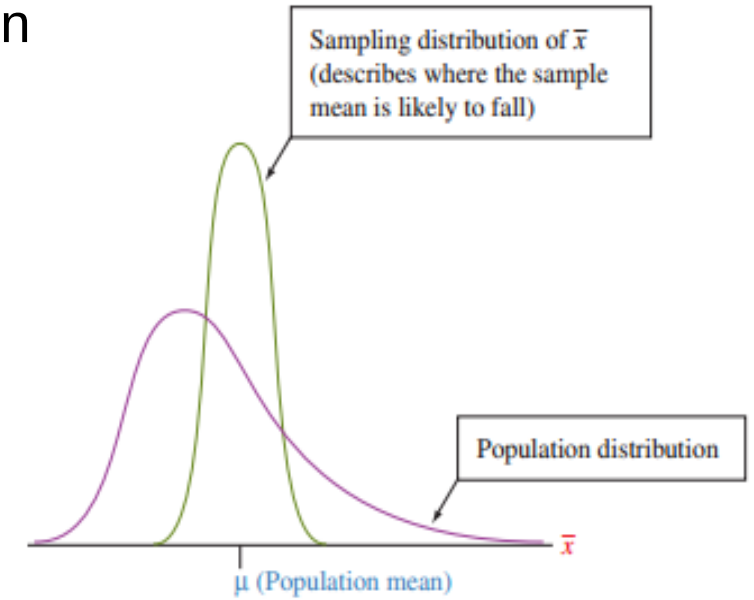
- It incorporates a margin of error which helps us to gauge the accuracy of the point estimate



Point Estimate and sample statistics (Interval Estimate)

Property 1: Good estimator has sampling distribution centered at the parameter An estimator with this property is unbiased.

Property 2: Good estimator has smaller standard error than other estimators



The Sample Mean $\overline{X}$Is an Unbiased Estimator

Central Limit Theorem

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

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

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

- 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 $SE = \sigma / \sqrt{n}$, 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.

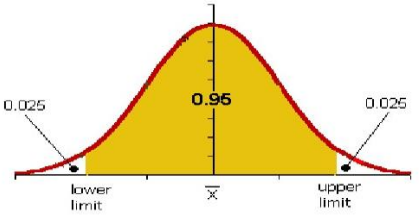
Sample size / skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.

- 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

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.

Confidence intervals

Confidence intervals



- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer and Chris Penny (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 2 \times SE$$

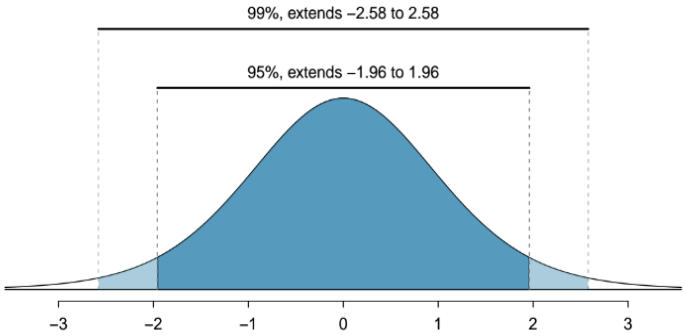
$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$$\begin{aligned} \bar{x} \pm 2 \times SE &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \end{aligned}$$

A more accurate interval

Confidence interval, a general formula

$$\textit{point estimate} \pm z^* \times SE$$



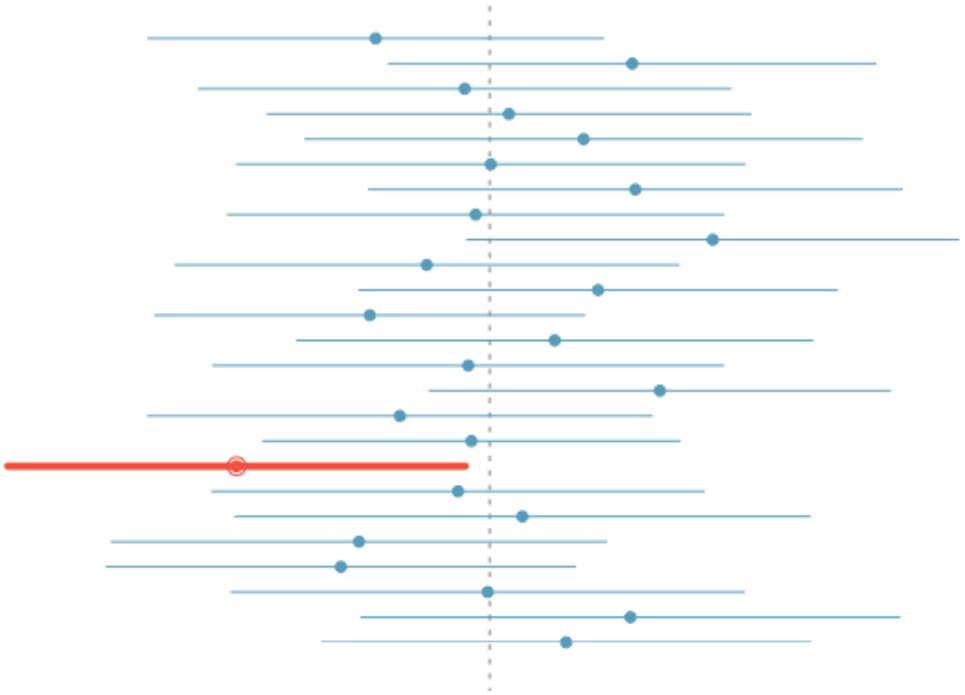
Conditions when the point estimate = $\bar{x}$

1. *Independence*: Observations in the sample must be independent
    - random sample/assignment
    - if sampling without replacement, n < 10% of population
1. *Sample size / skew*: $n \geq 30$ and population distribution should not be extremely skewed

Note: We will discuss working with samples where $n < 30$ in the next chapter.

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* ± *2 x SE*.
- Then about 95% of those intervals would contain the true population mean (μ).
- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.

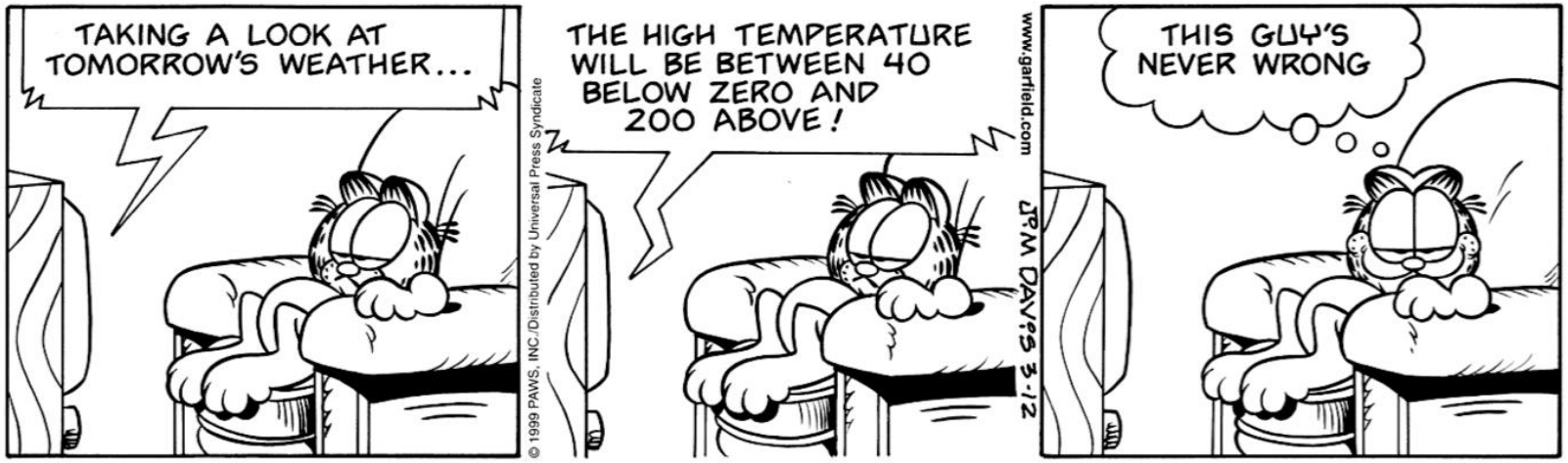


Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Changing the confidence level

$$\textit{point estimate} \pm z^* \times SE$$

- In a confidence interval, $z^* \times SE$ is called the *margin of error*, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust z^* in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^* = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate z^* for any confidence level.

Hypothesis testing

Remember when...

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

$\hat{p}_{\text{males}} = 21 / 24 = 0.88$

$\hat{p}_{\text{females}} = 14 / 24 = 0.58$

Possible explanations:

- Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance.
  → null (nothing is going on)
- Promotion and gender are *dependent*, there is gender discrimination, observed difference in proportions is not due to chance.
  → alternative (something is going on)

Recap: hypothesis testing framework

- 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 or traditional methods based on the central limit theorem (coming up next...).
- 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.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

Number of college applications

A similar survey asked how many colleges students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all Duke students apply to is higher than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

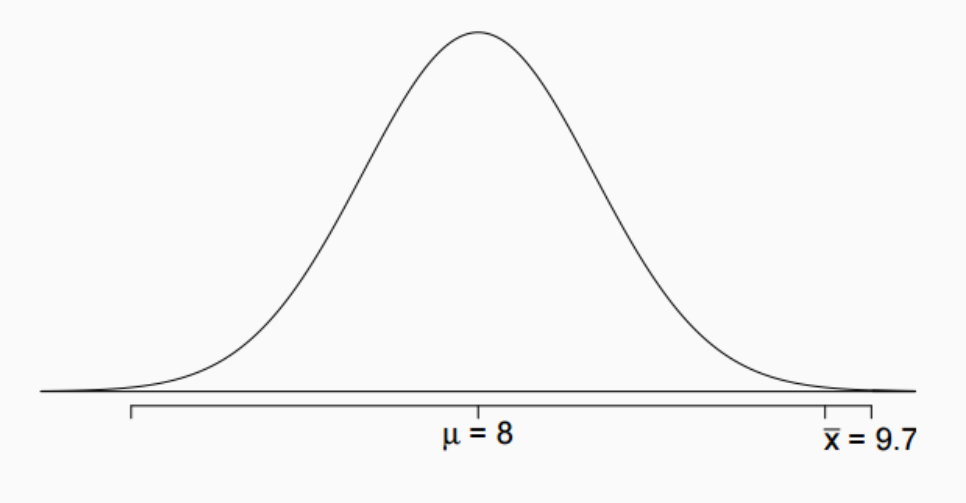
Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability
  - We start with the assumption the average number of colleges Duke students apply to is 8 (as recommended) $H_0 : \boldsymbol{\mu} = 8$
- We test the claim that the average number of colleges Duke students apply to is greater than 8

$$H_A : \boldsymbol{\mu} > 8$$

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*.



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

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

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*?

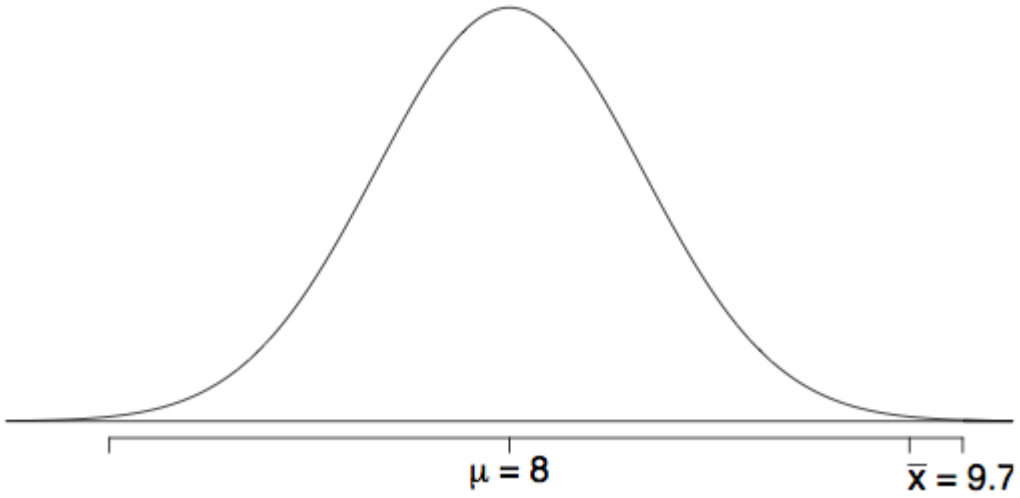
Yes, and we can quantify how unusual it is using a p-value.

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.
- 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.

Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



$$P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$$

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject* H_0.
- The data provide convincing evidence that Duke students apply to more than 8 schools on average.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is *not due to chance* or sampling variability.

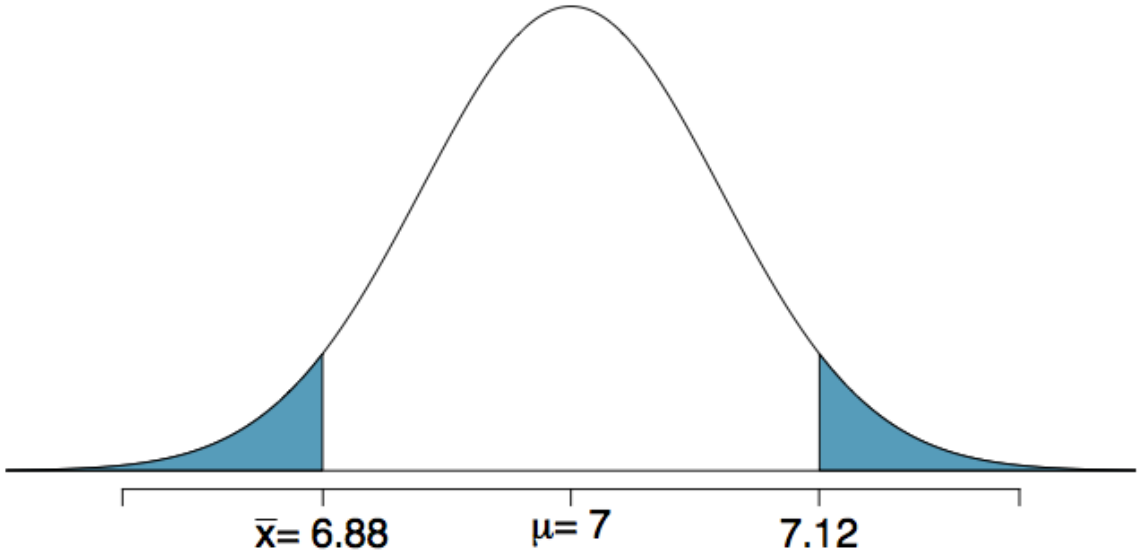
Two-sided hypothesis testing with p-values

- 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

$$H_0: \mu = 7$$

$$H_A: \mu \neq 7$$

- Hence the p-value would change as well:



p-value
$= 0.0485 \times 2$
$= 0.097$

Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

2. Hypotheses

Null: $H_0: p = p_0$, where p_0 is the hypothesized value.

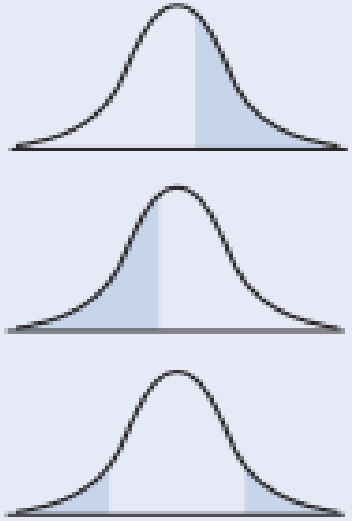
Alternative: $H_a: p \neq p_0$ (two-sided) or $H_a: p < p_0$ (one-sided) or $H_a: p > p_0$ (one-sided)

3. Test statistic

$$z = \frac{\hat{p} - p_0}{se_0} \text{ with } se_0 = \sqrt{p_0(1 - p_0)/n}$$

4. P-value

Alternative hypothesis	P-value
$H_a: p > p_0$	Right-tail probability
$H_a: p < p_0$	Left-tail probability
$H_a: p \neq p_0$	Two-tail probability



5. Conclusion

Smaller P-values give stronger evidence against H_0. If a decision is needed, reject H_0 if the P-value is less than or equal to the preselected significance level (such as 0.05). Relate the conclusion to the context of the study.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

Type 1 error

Possible Results of a Legal Trial

	Legal Decision	
Defendant	**Acquit**	**Convict**
Innocent (H_0)	Correct decision	Type I error
Guilty (H_a)	Type II error	Correct decision

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α -- increasing α increases the Type 1 error rate.

Choosing a significance level

- Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.
- We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

the next two slides provide a brief summary of hypothesis testing...

Recap: Hypothesis testing for a population mean

1. Set the hypotheses
    - H_0: μ = null value
    - H_A: μ < or > or ≠ null value
2. Calculate the point estimate
3. Check assumptions and conditions
    - Independence: random sample/assignment, 10% condition when sampling without replacement
    - Normality: nearly normal population or $n \geq 30$, no extreme skew -- or use the t distribution (Ch 5)
4. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}}$$

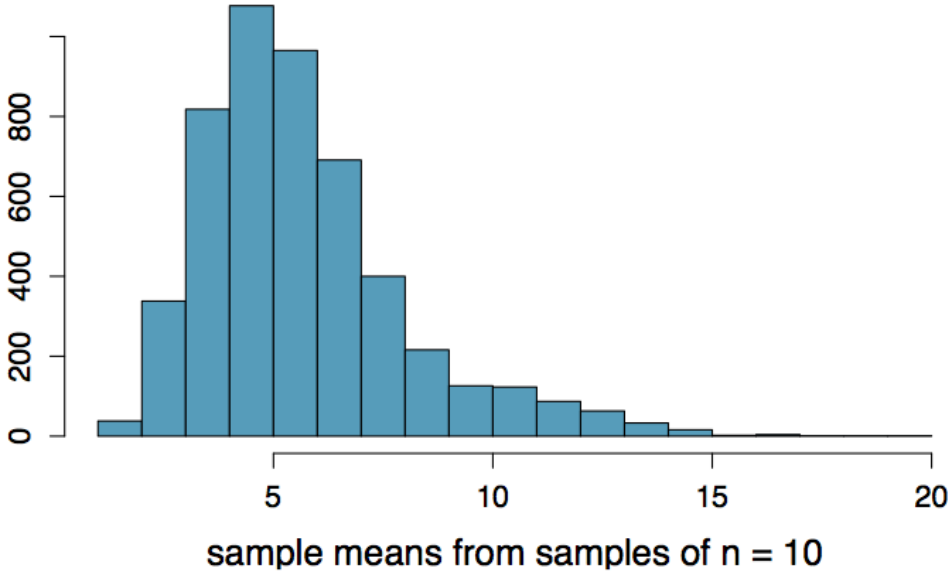
5. Make a decision, and interpret it in context

- If p-value < α, reject H_0, data provide evidence for H_A
- If p-value > α, do not reject H_0, data do not provide evidence for H_A

Examining the Central Limit Theorem

Average number of basketball games attended (cont.)

Sampling distribution, $n = 10$:



What does each observation in this distribution represent?

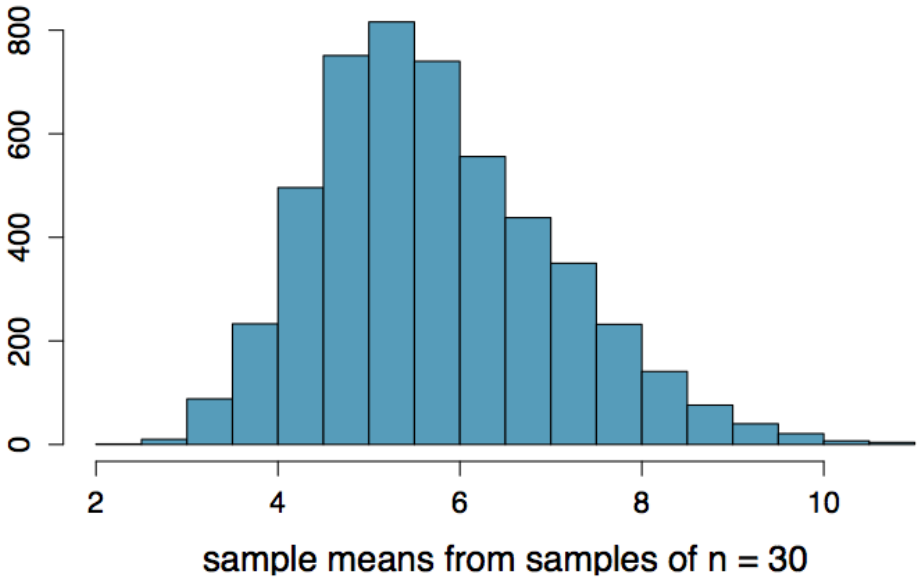
Sample mean ($\bar{x}$) of samples of size n = 10.

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Smaller, sample means will vary less than individual observations.

Average number of basketball games attended (cont.)

Sampling distribution, $n = 30$:



How did the shape, center, and spread of the sampling distribution change going from $n = 10$ to $n = 30$?

Shape is more symmetric, center is about the same, spread is smaller.

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.
- 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 these at the end of this section.

Confidence intervals for nearly normal point estimates

A confidence interval based on an unbiased and nearly normal point estimate is

$$\textit{point estimate} \pm z^* \, SE$$

where z^* is selected to correspond to the confidence level, and SE represents the standard error.

Remember that the value $z^* \, SE$ is called the *margin of error*.

Hypothesis testing for nearly normal point estimates

The third National Health and Nutrition Examination Survey collected body fat percentage (BF%) and gender data from 13,601 subjects ages 20 to 80. The average BF% for the 6,580 men in the sample was 23.9, and this value was 35.0 for the 7,021 women. The standard error for the difference between the average men and women BF%s was 0.114. Do these data provide convincing evidence that men and women have different average BF%s. You may assume that the distribution of the point estimate is nearly normal.

Hypothesis testing for nearly normal point estimates

The third National Health and Nutrition Examination Survey collected body fat percentage (BF%) and gender data from 13,601 subjects ages 20 to 80. The average BF% for the 6,580 men in the sample was 23.9, and this value was 35.0 for the 7,021 women. The standard error for the difference between the average men and women BF%s was 0.114. Do these data provide convincing evidence that men and women have different average BF%s. You may assume that the distribution of the point estimate is nearly normal.

1. Set hypotheses
2. Calculate point estimate
3. Check conditions
4. Draw sampling distribution, shade p-value
5. Calculate test statistics and p-value, make a decision

Hypothesis testing for nearly normal point estimates (cont.)

1. The null hypothesis is that men and women have equal average BF%, and the alternative is that these values are different.

$$H_0: \mu_{men} = \mu_{women} \qquad H_A: \mu_{men} \neq \mu_{women}$$

Hypothesis testing for nearly normal point estimates (cont.)

1. The null hypothesis is that men and women have equal average BF%, and the alternative is that these values are different.

$$H_0: \mu_{men} = \mu_{women} \qquad H_A: \mu_{men} \neq \mu_{women}$$

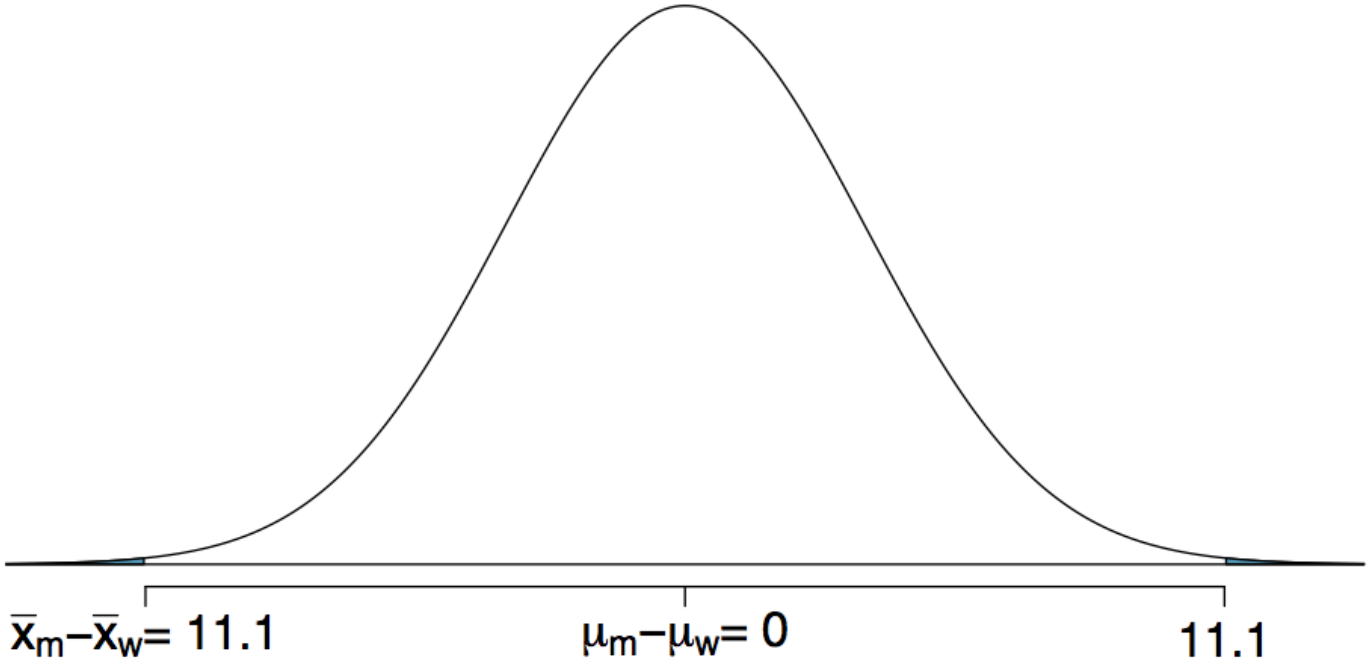
1. The parameter of interest is the average difference in the population means of BF%s for men and women, and the point estimate for this parameter is the difference between the two sample means:

$$\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35.0 = -11.1$$

1. We are assuming that the distribution of the point estimate is nearly normal (we will discuss details for checking this condition in the next chapter, however given the large sample sizes, the normality assumption doesn't seem

Hypothesis testing for nearly normal point estimates (cont.)

4. The sampling distribution will be centered at the null value ($\mu_{men} - \mu_{women} = 0$), and the p-value is the area beyond the observed difference in sample means in both tails (lower than -11.1 and higher than 11.1).



Hypothesis testing for nearly normal point estimates (cont.)

5. The test statistic is computed as the difference between the point estimate and the null value (-11.1 - 0 = -11.1), scaled by the standard error.

$$Z = \frac{11.1 - 0}{0.114} = 97.36$$

The Z score is huge! And hence the p-value will be tiny, allowing us to reject H_0 in favor of H_A.

These data provide convincing evidence that the average BF% of men and women are different.

Non-normal point estimates

- We may apply the ideas of confidence intervals and 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 confidence intervals and 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:

1. Learn new methods that are appropriate for the data.
2. Consult a statistician.
3. ~~Ignore the failure of conditions.~~ This last option effectively invalidates any analysis and may discredit novel and interesting findings.

Sample Size and Power

Finding a sample size for a certain margin of error

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

We know that the critical value associated with the 96% confidence level:
$z^* = 2.05$.

$$4 \geq 2.05 * 18/\sqrt{n} \rightarrow n \geq (2.05 * 18/4)^2 = 85.1$$

The minimum number of children required to attain the desired margin of error is 85.1. Since we can't sample 0.1 of a child, we must sample at least 86 children (round up, since rounding down to 85 would yield a slightly larger margin of error than desired).

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592

Hypothesis testing possibilities

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1 - \beta$

Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)

Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)

Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is 1 - β

In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

Type 2 error rate

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

- The answer is not obvious.
- If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- Clearly, β depends on the effect size (δ)

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is *greater* than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

H_0: $\mu = 130$

H_A: $\mu > 130$

We'll start with a very specific question -- “What is the power of this hypothesis test to correctly detect an *increase* of 2 mmHg in average blood pressure?”

Calculating power

The preceding question can be rephrased as “How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?”

Hint: Break this down into two simpler problems

- Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
- Problem 2: What is the probability that we would reject H_0 if $\bar{x}$ had come from N(mean = 132, SE = 25 / $\sqrt{100}$ = 2.5), i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?

Determine how power changes as sample size, standard deviation of the sample, α, and effect size increases.

Problem 1

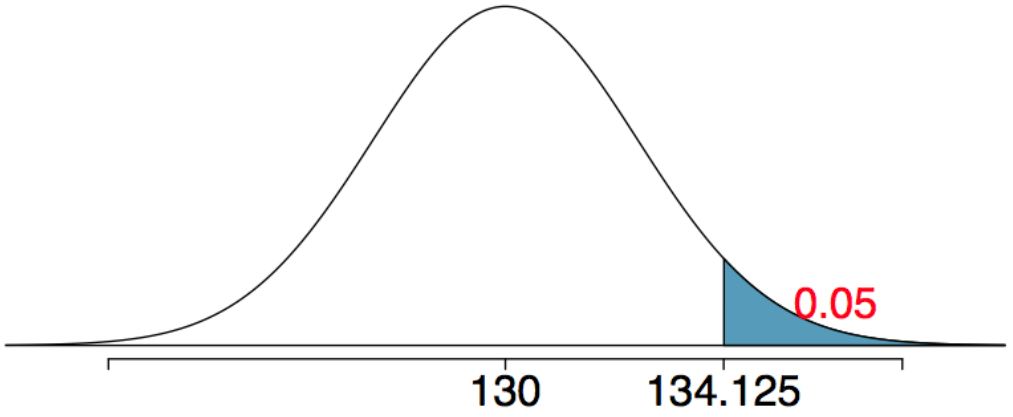
Which values of $\bar{x}$ represent sufficient evidence to reject H_0? (Remember H_0: $\mu = 130$, H_A: $\mu > 130$)

$$P(Z > z) < 0.05 \quad \Rightarrow \quad z > 1.65$$

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} > 1.65$$

$$\bar{x} > 130 + 1.65 \times 2.5$$

$$\bar{x} > 134.125$$



Any $\bar{x} > 134.125$ would be sufficient to reject H_0 at the 5% significance level.

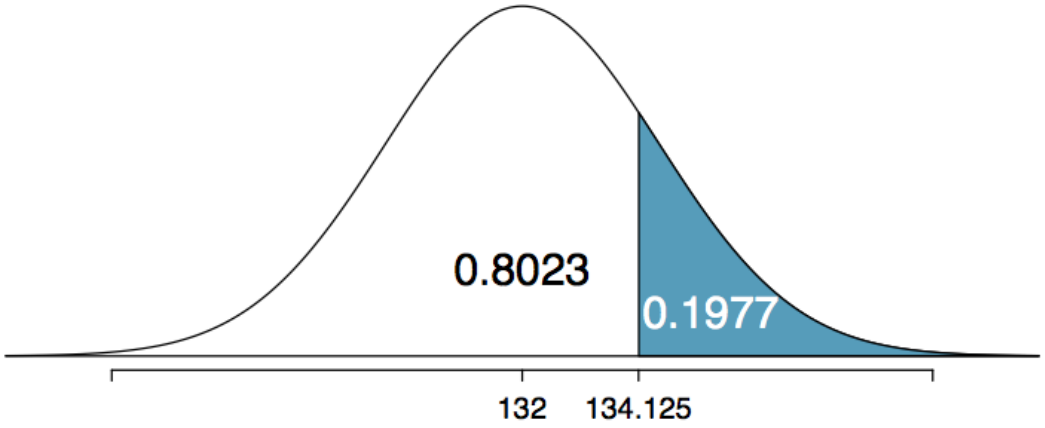
Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ did come from N(mean = 132, SE = 2.5).

This is the same as finding the area above $\bar{x}$ = 134.125 if $\bar{x}$ came from N(132, 2.5).

$$Z = \frac{134.125 - 132}{2.5}$$
$$= 0.85$$

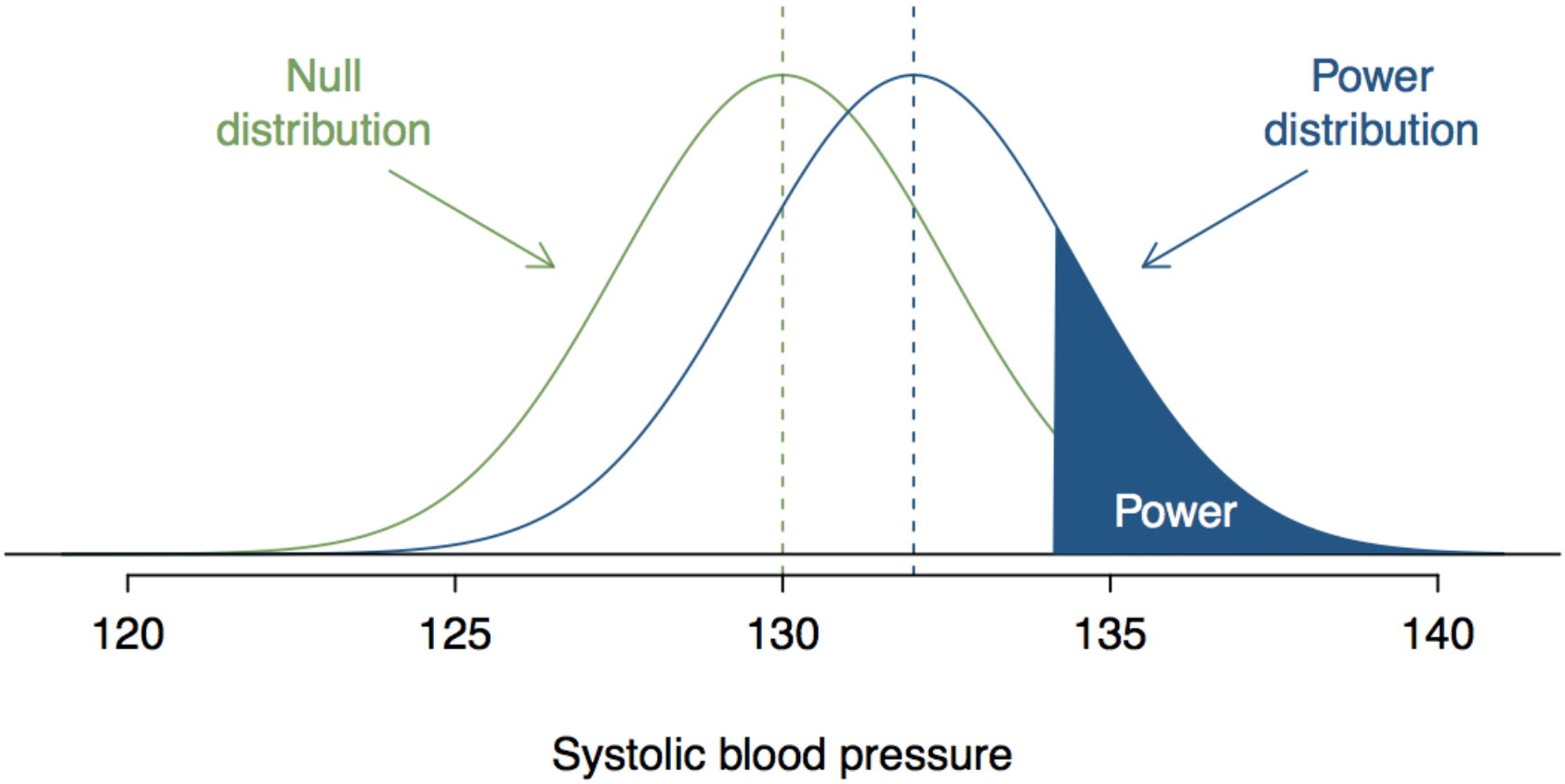
$$P(Z > 0.85) = 1 - 0.8023$$
$$= 0.1977$$



The probability of rejecting H_0: μ = 130, if the true average systolic blood pressure of employees at this company is 132 mmHg, is 0.1977 which is the power of this test.

Therefore, β = 0.8023 for this test.

Putting it all together



Achieving desired power

There are several ways to increase power (and hence decrease the Type 2 Error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller *s* we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase α, which will make it more likely to reject H_0 (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.

Recap - Calculating Power

Begin by picking a meaningful effect size δ and a significance level α.

Calculate the range of values for the point estimate beyond which you would reject H_0 at the chosen α level.

Calculate the probability of observing a value from preceding step if the sample was derived from a population where $\bar{x} \sim N(\mu_{H0} + \delta, SE)$.

Statistical vs Practical Significance

All else held equal, will the p-value be lower if $n = 100$ or $n = 10{,}000$?

a) $n = 100$

b) $n = 10{,}000$

All else held equal, will the p-value be lower if $n = 100$ or $n = 10{,}000$?

a) $n = 100$

b) $n = 10{,}000$

Suppose: $\bar{x} = 50$, $s = 2$, H_0: $\mu = 49.5$, H_A: $\mu > 49.5$

$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5, \quad p\text{-value} = 0.0062$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25, \quad p\text{-value} \approx 0$$

As n increases: $SE \downarrow$, $Z \uparrow$, p-value $\downarrow$

Test the hypothesis H_0: $\mu = 10$ vs. H_A: $\mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$			
$n = 5000$			

- When σ is known:
  Test statistic

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- When σ is unknown:
  Test statistic

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Test the hypothesis H_0: $\mu = 10$ vs. H_A: $\mu > 10$ for the following 8 samples. Assume $\sigma = 2$.

$\bar{x}$	10.05	10.1	10.2
$n = 30$	$p - value = 0.45$	$p - value = 0.39$	$p - value = 0.29$
$n = 5000$	$p - value = 0.04$	$p - value = 0.0002$	$p - value \approx 0$

When n is large, even small deviations from the null (small effect sizes), which may be considered practically insignificant, can yield statistically significant results.

Statistical vs Practical Significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (*effect size*), even when the difference is not practically significant.
- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real, but also large enough to matter).
- The role of a statistician is not just in the analysis of data, but also in planning and design of a study.

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

- R.A. Fisher