Lecture 16
Last time

- The standard normal distribution
- Computing probabilities
- Sums of independent normals
Histogram of midterm 1 scores

Frequency

scores out of 50

15 20 25 30 35 40 45 50

0
5
10
15
Midterm 1 results

• 5-number summary (19, 35, 40, 45, 49)

• If you did somewhere close to the median, you’re in good shape

• If you are far below the first quartile, you should see me or one of the TAs
Toy examples

Let’s conduct a simple experiment

• Maybe we’re planting potatoes using a new fertilizer and the response is the weight of the potato when harvested

• Or, we’re testing a new drug designed to lower cholesterol levels for a certain group of patients

• Or, we’re conducting a survey of adult males living in the U.S. and asking each respondent to give their height in inches
Toys

• In each case we perform some kind of random experiment \( n \) times and then measure something about the outcome

• We let \( X_1, X_2, \ldots, X_n \) denote random variables that represent the results of our experiment
Toys

• Formally, we assume that each trial is independent of the others, and that each random variable $X_i$ has the same probability distribution.

• In each case, we are interested in the population mean; the average weight of a potato using the new fertilizer, the average cholesterol level of a patient using the new drug, and so on.
Toys

• We will estimate the population mean using the average of our random variables

\[ \bar{X} = \frac{1}{n} \sum_{i} X_i \]
Suppose each random variable has this probability density function (can this happen?)
The population mean is 6.5
Toys

• Suppose we conduct 10 trials (10 potato plants, 10 patients, 10 randomly selected adult males)

3.1, 4.5, 13.4, 1.1, 1.9, 1.6, 5.6, 1.9, 3.1, 0.8

... and again

2.6, 8.8, 2.2, 16.1, 15.3, 11.0, 0.7, 1.5, 1.8, 14.9
Toys

- In the first case, the sample mean is 3.7 and in the second case, it is 6.7

- Notice that the sample mean is itself a random variable; each time we conduct a new set of 10 experiments, we get a different sample mean
1,000 sample means (n=10 for each)
1,000 sample means (n=100 for each)
1,000 sample means (n=1,000 for each)
Some things to notice

• The sample means tend to be distributed around the population mean

• As we increase the number of trials we run, the variation in the sample mean decreases
Some facts

• The expected value of the sample mean

\[ E\bar{X} = \frac{1}{n} \sum_i E X_i = \frac{1}{n} \sum_i \mu = \mu \]
Some facts

- The variance of the sample mean (recall the experiments are independent)

\[
\text{var} \bar{X} = \frac{1}{n^2} \sum_i \text{var} X_i = \frac{1}{n^2} \sum_i \sigma^2 = \sigma^2 / n
\]
Some things to notice

• The sample means tend to be distributed around the population mean

• As we increase the number of trials we run, the variation in the sample mean decreases

• Even though the distribution of each random variable is not normal, the distribution of the sample means takes on our friendly bell shape; this is called the *central limit effect*
Parameters and estimates

- A parameter is a numerical characteristic of a distribution

- In our experiments, we are after the population mean

- Binomial: The number of trials $n$ and the success probability $p$ are parameters

- Normal: The mean and variance are parameters
Parameters and estimates

- We use data to produce an *estimate* of an unknown parameter

- Data often consist of a sample drawn from a probability distribution
Standard errors

- Given the sample mean, can we estimate how close we are to the population mean?
- The *standard error* is an estimate of the \( \text{sd}(X) \)

\[
\text{Sample standard deviation} = \frac{\text{Sample size}}{\sqrt{\text{Sample size}}}
\]
Lecture 17
Last time

• Some toy examples

• Properties of the sample mean

• The central limit effect
The toy examples

• In each case we performed some kind of random experiment $n$ times and then measured something about the outcome

• Our interest was in estimating the population mean $\mu$

• Formally, we let $X_1, X_2, ..., X_n$ denote random variables representing the results of our experiments; we assumed that each trial was independent
The examples

- We studied the sample mean

\[ \bar{X} = \frac{1}{n} \sum_{i} X_i \]

- This is an \textit{estimate} of the population mean \( \mu \)

- We can also think of it as a random variable; each time we repeat our experiment we get a different sample mean
We observed...

- The sample means tend to be distributed around the population mean $\mu$

- As we increase the number of trials we run, the variation in the sample mean decreases
Some facts

• The expected value of the sample mean

\[ E\bar{X} = \frac{1}{n} \sum_{i} EX_i = \frac{1}{n} \sum_{i} \mu = \mu \]

• The variance of the sample mean

\[ \text{var}\bar{X} = \frac{1}{n^2} \sum_{i} \text{var} X_i = \frac{1}{n^2} \sum_{i} \sigma^2 = \sigma^2 / n \]
The central limit effect

- Even though the distribution of each random variable is not normal, the distribution of the sample means takes on our friendly bell shape.
To remember

• The sample mean always has an expected value $\mu$ and its standard deviation is $\sigma/\sqrt{n}$ provided the experiments are independent.

• The central limit effect holds regardless of the shape of the distribution of our random variables $X_i$. 
In reality...

- When we conduct an experiment, we want to learn something about the population mean.

- Obviously, we don’t know \( \mu \) (or we wouldn’t need to run an experiment) and we don’t know the population variance \( \sigma^2 \).

- After we run an experiment we get the sample mean; what does that tell us about \( \mu \)?
In reality...

• Suppose I run an experiment and get the following 100 numbers

0.97 2.71 1.81 0.96 4.03 5.21 2.38 2.39 0.83 2.62 4.30 1.14 1.85 2.48 2.28
2.19 2.45 4.88 5.09 1.55 2.52 3.09 5.46 1.61 2.10 3.79 2.38 2.98 5.68 3.34
3.15 2.11 2.31 4.58 5.36 4.25 1.07 2.21 3.46 2.10 1.15 1.88 1.84 4.77 2.18
2.22 3.64 3.51 4.82 4.39 2.09 1.14 4.80 2.18 6.44 1.91 1.03 0.88 2.69 0.29
1.01 3.57 2.96 2.83 1.94 2.13 2.30 7.75 3.39 1.88 2.85 2.89 4.21 2.41 3.49
5.94 3.25 5.17 3.75 2.13 0.53 7.47 3.49 3.44 0.95 2.12 3.20 2.72 2.56 1.66
4.23 2.87 6.61 5.53 4.29 2.59 3.45 2.05 1.90 3.46

• The sample mean is 3.0, what can we say about the population mean?
Lecture 18
Last time

- Homework problems (gout, wolves, the usual)
- The central limit effect and confidence intervals
Simple setup

\[ \bar{X} = \frac{1}{n} \sum_{i} X_i \]
Some facts

• The sample mean always has an expected value \( \mu \) and its standard deviation is \( \sigma / \sqrt{n} \) provided the experiments are independent.

• The central limit effect holds regardless of the shape of the distribution of our random variables \( X_i \).
The central limit effect

- Even though the distribution of each random variable is not normal, the distribution of the sample means takes on our friendly bell shape.
In reality...

• When we conduct an experiment, we want to learn something about the population mean

• Obviously, we don’t know $\mu$ (or we wouldn’t need to run an experiment) and we don’t know the population variance $\sigma^2$

• After we run an experiment we get the sample mean; what does that tell us about $\mu$ ?
In reality...

• Suppose I run an experiment and get the following 100 numbers

0.97 2.71 1.81 0.96 4.03 5.21 2.38 2.39 0.83 2.62 4.30 1.14 1.85 2.48 2.28
2.19 2.45 4.88 5.09 1.55 2.52 3.09 5.46 1.61 2.10 3.79 2.38 2.98 5.68 3.34
3.15 2.11 2.31 4.58 5.36 4.25 1.07 2.21 3.46 2.10 1.15 1.88 1.84 4.77 2.18
2.22 3.64 3.51 4.82 4.39 2.09 1.14 4.80 2.18 6.44 1.91 1.03 0.88 2.69 0.29
1.01 3.57 2.96 2.83 1.94 2.13 2.30 7.75 3.39 1.88 2.85 2.89 4.21 2.41 3.49
5.94 3.25 5.17 3.75 2.13 0.53 7.47 3.49 3.44 0.95 2.12 3.20 2.72 2.56 1.66
4.23 2.87 6.61 5.53 4.29 2.59 3.45 2.05 1.90 3.46
Not very normal looking.
In reality...

• Can the population mean be 4? 2? 25?

• We can use the central limit effect to our advantage:

  We know that the sample mean has an (approximately) normal distribution with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$

  If we knew $\sigma^2$ we could standardize $\bar{X}$!
In reality...

• To be precise, if we have a reasonably large sample size $n$,

\[
\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}
\]

has an approximately standard normal distribution
In reality...

• Putting this fact to good use, recall the 95% rule: for a standard normal distribution

\[ \text{pr} \left( -2 < Z < 2 \right) = 0.954 \]
In reality...

• Or, using what we know about the sample mean

$$\text{pr} \left( -2 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 2 \right) = 0.954$$
In reality...

- We can then turn this around so that

$$\Pr\left(\bar{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.954$$
In reality...

- We refer to this as a confidence interval for $\mu$

\[ \Pr \left( \bar{X} - 2\sigma / \sqrt{n} < \mu < \bar{X} + 2\sigma / \sqrt{n} \right) = 0.954 \]

- It means that if we perform our experiment lots and lots of times, the chance that our population mean is within the indicated interval is 0.954
Repeating the experiment 10 times
Repeating the experiment 1,000 times
Repeating the experiment 1,000 times
Interpretation!

• If we repeat our experiment multiple times, we would see that our population mean is contained in our confidence interval 95% of the time
In reality...

• Of course, we don’t really know $\sigma^2$

• Instead, we can estimate it using the sample standard deviation

• In our case the sample standard deviation is 1.9
Standard errors

- The *standard error* is an estimate of the $\text{sd}(\bar{X}) = \sigma / \sqrt{n}$

\[
\text{Sample standard deviation} = \frac{\sigma}{\sqrt{n}}
\]
In reality...

- It turns out that for large samples, we can use

\[
\frac{\bar{X} - \mu}{\text{se}(\bar{X})}
\]

to form our confidence intervals, or rather,

\[
(\bar{X} - 2\text{se}(\bar{X}), \bar{X} + 2\text{se}(\bar{X}))
\]
In reality...

- In our case, the standard error is

\[ \frac{1.9}{\sqrt{100}} = 0.19 \]

- and our confidence interval is

\[ (3 - 2 \times 0.19, 3 + 2 \times 0.19) = (2.62, 3.38) \]
Parameters and estimates

• Data often consist of independent samples drawn from a fixed probability distribution

• A parameter is a numerical characteristic of the probability distribution, generically denoted $\theta$

• We use data to produce an estimate of an unknown parameter, generically denoted $\hat{\theta}$

• Statistical inference is the process of using samples of data to make useful statements about an unknown parameter $\theta$
Parameters and estimates

- *Point estimates* are those which refer to single parameters

- We distinguish them from *interval estimates* like confidence intervals
Some concepts

Given a parameter $\theta$ and an estimate $\hat{\theta}$

- **Bias**: An estimate is unbiased if $E\hat{\theta} = \theta$

- **Precision**: How variable is the estimator if we repeatedly sample new data?
Comparing means

• Suppose we have the results of two experiments

• Revisiting our toy examples, we might want to compare two fertilizers and their effect on the yield of potato plants; or we might compare a new cholesterol drug to some standard treatment
What can we say about the difference?
What can we say about the difference?