Monday, February 11: Interval estimates

Estimation of parameters

A common problem in statistics: we are given a sample (an independent and identically distributed set of random variables $X_1, X_2, \ldots, X_n$). Each individual distribution is known to be from a particular family but one of its parameters is unknown. We observe these random variables once and obtain values $x_1, x_2, \ldots, x_n$. What do these observations tell us about the value of the unknown parameter in the distribution of $X$?

One thing we might be able to do is to compute a confidence interval for the parameter, an idea that we now explore.
Monday, February 11: Interval estimates

A random interval is an interval of $\mathbb{R}$ whose endpoints are random variables

**Example** Let $X \sim \chi^2(16)$. Then the interval $(X, 3.3X)$ is a random interval. What is the probability that $26.3 \in (X, 3.3X)$? Also, what is the expected value of the length of the interval?
Monday, February 11: Interval estimates

We have that $X < 26.3 < 3.3X$ if and only if $X < 26.3$ and $X > 26.3/3.3 \approx 7.97$.

If $F(x)$ is the cumulative distribution function for a $\chi^2(16)$ distribution, then the probability is $F(26.3) - F(26.3/3.3)$, which is 0.8998.

For the expected value of the length, we have

$$E(3.3X - X) = E(2.3X) = (2.3)(16) = 36.8.$$
Monday, February 11: Interval estimates

**Example** Suppose that \( \bar{X} \) is the sample mean of an independent and identically distributed set \( X_1, \ldots, X_n \), each with distribution \( N(\mu, \sigma) \), where \( \sigma \) is known but \( \mu \) is unknown.

What is

\[
P \left( \mu \in \left( \bar{X} - 2 \frac{\sigma}{\sqrt{n}}, \bar{X} + 2 \frac{\sigma}{\sqrt{n}} \right) \right) \]
Monday, February 11: Interval estimates

By the smaller version of the central limit theorem,

\[ \bar{X} \sim N(\mu, \sigma/\sqrt{n}), \]

which means that

\[ Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \]

This tells us that if \( F(z) \) is the cumulative distribution function of a standard normal distribution, then

\[ P(-2 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 2) = F(2) - F(-2) = 0.9544997, \]

which we could nearly guess from the 66-95-99.7 Rule.
Monday, February 11: Interval estimates

Solving the above gives that

\[
P \left( \mu \in \left( X - 2 \frac{\sigma}{\sqrt{n}}, X + 2 \frac{\sigma}{\sqrt{n}} \right) \right) = F(2) - F(-2) = 0.9544997.
\]

Note that the original question really boils down to the question: what is the probability that the normally distributed random variable \( X \) will lie within 2 of its standard deviations from its mean?
Definition A 95% confidence random interval for a parameter $\theta$ is a random interval $J$ with the property that $P(\theta \in J) = 0.95$. In this context, 0.95 is called the confidence level (and it need not be 0.95, although that is most common).

A 95% confidence interval for a parameter $\theta$ is an observation of a 95% confidence random interval for a parameter $\theta$.

Let’s reinterpret the example from yesterday...
Example Suppose that $\overline{X}$ is the sample mean of an independent and identically distributed set $X_1, \ldots, X_n$, each with distribution $N(\mu, \sigma)$, where $\sigma$ is known but $\mu$ is unknown.

To find a 95% confidence random interval for $\mu$, find $z^*$ with the property that $F(z^*) - F(-z^*) = 0.95$, where $F(x)$ is the cumulative distribution function of a standard normal distribution.

Then

$$P \left( \mu \in \left( \overline{X} - z^* \frac{\sigma}{\sqrt{n}}, \overline{X} + z^* \frac{\sigma}{\sqrt{n}} \right) \right) = F(z^*) - F(-z^*) = 0.95,$$

so

$$\left( \overline{X} - z^* \frac{\sigma}{\sqrt{n}}, \overline{X} + z^* \frac{\sigma}{\sqrt{n}} \right)$$

is a 95% confidence random interval for $\mu$. 

Tuesday, February 12: Interval estimates
Note that this is a 95% confidence random interval for $\mu$, not the 95% confidence random interval for $\mu$

There are many (infinitely many) 95% confidence random intervals for $\mu$; this is simply one of them

Also note that it is always for a parameter — it must be a confidence random interval for something, not just a confidence random interval
Tuesday, February 12: Interval estimates

Where do we find $z^*$? Drawing the picture of a standard normal probability density function and shading the desired area that should be 0.95, we find that if $Q(p)$ is the quantile function of a standard normal distribution, then

$$z^* = Q(0.975) \approx 1.959964.$$

This $z^*$ is called (in my own terminology) the 0.95 central quantile for a standard normal distribution.
Tuesday, February 12: Interval estimates

In real-world applications, it is unlikely that we would ever know $\sigma$ without knowing $\mu$ (why?).

What might we use instead of $\sigma$ to obtain a confidence random interval for $\mu$?
Tuesday, February 12: Interval estimates

How about $S^2$? Well, we have discussed that

$$\frac{X - \mu}{\sigma / \sqrt{n}} \sim N(0, 1) \quad \text{and} \quad nS^2 / \sigma^2 \sim \chi^2(n - 1),$$

and that these two are independent (which may be the hardest part to show)

This means that

$$T = \frac{(X - \mu) / (\sigma / \sqrt{n})}{\sqrt{nS^2 / ((n - 1)\sigma^2)}} \sim t(n - 1).$$

Simplifying gives that

$$T = \frac{X - \mu}{S / \sqrt{n - 1}} \sim t(n - 1).$$
Tuesday, February 12: Interval estimates

Just as $Z = (X - \mu) / \sigma$ is the *standardized* version of $X \sim N(\mu, \sigma)$, the random variable

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n - 1}}$$

is called the *studentized* version of $X$

The standardized version of $X$ (with $\sigma$) has a standard normal distribution

The studentized version of $X$ has a $t(n - 1)$ distribution
Tuesday, February 12: Interval estimates

**Example** Suppose that $\bar{X}$ is the sample mean of an independent and identically distributed set $X_1, \ldots, X_n$, each with distribution $N(\mu, \sigma)$, where $\sigma$ and $\mu$ are unknown.

To find a 95% confidence interval for $\mu$, find $t^*$ with the property that $F(t^*) - F(-t^*) = 0.95$, where $F(x)$ is the cumulative distribution function of a $t(n-1)$ distribution.

Then

$$P \left( \mu \in \left( \bar{X} - t^* \frac{S}{\sqrt{n-1}}, \bar{X} + t^* \frac{S}{\sqrt{n-1}} \right) \right) = F(t^*) - F(-t^*) = 0.95,$$

so

$$\left( \bar{X} - t^* \frac{S}{\sqrt{n-1}}, \bar{X} + t^* \frac{S}{\sqrt{n-1}} \right)$$

is a 95% confidence interval for $\mu$. 
Tuesday, February 12: Interval estimates

This $t^*$ is called (in my own terminology) the 0.95 central quantile for a $t(n - 1)$ distribution.
Wednesday, February 13: Interval estimates

What if we want to compare two means? One way is to find a confidence random interval for the difference between the two: let

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent samples,
- the distribution of each $X_i$ be $N(\mu_1, \sigma)$ and of each $Y_i$ be $N(\mu_2, \sigma)$ (with the same $\sigma$),
- the sample means be $\bar{X}$ and $\bar{Y}$, and the sample variances be $S_1^2$ and $S_2^2$.

Note that the four random variables in the last item are independent.

What can we say about the distribution of $\bar{Y} - \bar{X}$?
Wednesday, February 13: Interval estimates

By various previous results, we know that

\[
Y - X \sim N(\mu_2 - \mu_1, \sqrt{\sigma^2 / n + \sigma^2 / m})
\]

This means that

\[
\frac{(Y - X) - (\mu_2 - \mu_1)}{\sqrt{\sigma^2 / n + \sigma^2 / m}} \sim N(0, 1).
\]
Wednesday, February 13: Interval estimates

Also, $nS^2_1/\sigma^2 \sim \chi^2(n - 1)$ and $mS^2_2/\sigma^2 \sim \chi^2(m - 1)$, and they are independent (which needs checking), so their sum satisfies

\[
\frac{nS^2_1 + mS^2_2}{\sigma^2} \sim \chi^2(n + m - 2)
\]

This means that

\[
T = \frac{(\bar{Y} - \bar{X}) - (\mu_2 - \mu_1)}{\sqrt{\frac{nS^2_1 + mS^2_2}{n + m - 2} \left( \frac{1}{n} + \frac{1}{m} \right)}} \sim t(n + m - 2).
\]

The denominator is called the standard error, a term that we will define later:

\[
SE \sim \sqrt{\frac{nS^2_1 + mS^2_2}{n + m - 2} \left( \frac{1}{n} + \frac{1}{m} \right)}.
\]
Wednesday, February 13: Interval estimates

As per the usual method, we find that a 95% confidence random interval for $\mu_2 - \mu_1$ is

$$(\overline{Y} - \overline{X}) - t^*SE, (\overline{Y} - \overline{X}) + t^*SE),$$

where $t^*$ is the 0.95 central quantile of a $t(n + m - 2)$ distribution.