Chapter 32

Estimators

The term *estimator* is a fancy way of talking about functions computed from samples. It's basically probability theory applied to the specific task of making inferences about populations using sample data. This chapter will form the foundation for other topics we'll cover in this book: hypothesis testing (Chapter X), parametric models (Chapter Y), as well as linear models (Chapter Z). The math you'll need is no more advanced than what you already know. Indeed, all the estimator calculations and formulas were already introduced in the PROB chapters. How hard could this be?

32.1 Definitions

Let's start with some definitions.

32.1.1 General estimators concepts

We'll first review the general concepts that we presented in the introduction. The starting point is some *population* $\{x_1, x_2, x_3, x_4, x_5, \ldots, x_N\}$ that represents every item, person, animal, or event in a group of interest. Depending on the characteristics of the population, we'll build a *population model* $X \sim \textbf{model}(\theta)$, which is a probability distribution of a random variable that describes the population. The *model parameters* θ represent the unknown values of the "control knobs" of the probability distribution that describe this specific population.

For the probability calculations to make sense in this chapter (and most parts of the subsequent chapters), samples must be collected from the population using a method called *simple random sampling*. Random sampling is an *unbiased* technique of collecting a sample in such a way that every member of the population has an equal chance

of being included. A particular sample of size n from the population is denoted $\mathbf{x} \equiv \{x_1, x_2, \ldots, x_n\}$. The sample size n is usually much smaller than the population size N. We estimate properties of the population by computing the estimator value $f(\mathbf{x})$ from a particular sample $\mathbf{x}$.

We can think about estimators in two ways: practically and theoretically. Consider the set of functions f that take samples of size n as inputs,

$$f : \underbrace{\mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X}}_{n \text{ copies}} \quad \rightarrow \quad \mathbb{R},$$

where $\mathcal{X}$ denotes the type of values that you can encounter in the population, $f(\mathbf{x})$ denotes the value of the estimator computed on a particular sample $\mathbf{x} \equiv \{x_1, x_2, \ldots, x_n\}$, and $f(\mathbf{X})$ the value of the estimators computed on a *random sample* of the form $\mathbf{X} \equiv \{X_1, X_2, \ldots, X_n\}$, where each X_i is randomly selected value from the population.

In the practical context, the output value of the estimator is a number $f(\mathbf{x})$ that depends on a particular sample $\mathbf{x}$. However, because the sample is random, so too is the output value of the estimator. We can't predict the exact values that will end up in the sample, so we can't predict the exact value that's outputted by the estimator. The output value of the estimator $f(\mathbf{X})$ is therefore a random variable. It depends on the sample size n and the population. The value of the estimator computed from a random sample $f(\mathbf{X})$ allows us to characterize the variability of the estimates we are likely to observe.

32.1.2 Estimating a population parameter using a sample

Consider a particular sample $\mathbf{x} \equiv \{x_1, x_2, \ldots, x_n\}$ collected from the population using random sampling. Figure 32.1 shows the flow diagram for the general process of obtaining samples and computing estimator values $f(\mathbf{x})$ from the sample.

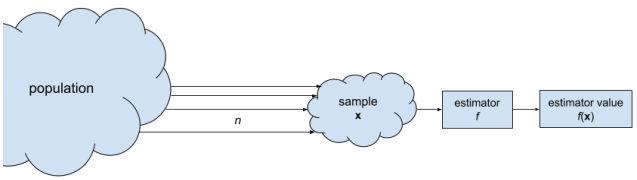

Figure 32.1: Using estimators in the context of statistical analysis of particular sample. The sample $\mathbf{x}$ is drawn from the population using the process of random selection. The estimator value $f(\mathbf{x})$ is a real number computed from the sample $\mathbf{x}$.

Let's look at the different types of estimates (estimator values) that we can compute from a particular sample.

- *Statistic*: any quantity computed from a sample. A statistic is a number that describes some characteristic of a sample.
- *Estimator value*: A statistic computed for the purpose of making an inference about a population. We can further subdivide the estimator values as follows:
  - ▻ *Parameter estimate* $\hat{\theta}$: a particular type of estimator used to compute estimates of population parameters, $\hat{\theta} \equiv f(\mathbf{x})$. For example, the sample mean estimator $\overline{x} = g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \mathbf{Mean}(\mathbf{x})$ is an estimator of the population mean μ.
  - ▻ *Test Statistic* $t, z, \chi^2, \ldots$: a particular type of statistic used as part of a hypothesis testing procedure. The value of the test statistic computed from a particular sample is used as input to a "decision rule" for reaching a conclusions about one of two competing statistical hypotheses.
  - ▻ *Confidence interval* $CI_{1-\alpha}$: A confidence interval is calculated from sample data and is used to estimate a parameter. The confidence interval will will contain the true population parameter in α% of cases if the process of collecting sample data and calculating the confidence interval is repeated a number of times.
  - ▻ *Estimated standard error of the estimator* $\hat{\theta}$ ($\hat{\mathbf{se}}_{\hat{\theta}}$): describes the variability of the estimates $\hat{\theta}$ we are likely to obtain for different random samples. For example estimated standard error of the sample mean is given by the formula $\mathbf{se}_{\overline{x}} \equiv \frac{s}{\sqrt{n}}$, where s is the standard deviation of the sample $\mathbf{x}$ (the square root of the sample variance $s^2 = h(\mathbf{x})$. The *estimated standard error* value $\hat{\mathbf{se}}_{\hat{\theta}}$ is an approximation for the true *standard error* value $\mathbf{se}_{\hat{\theta}}$, which will be defined in the next section.

Examples The following list shows some common estimates that we can compute on a particular sample of size n:

- Sample mean: $\overline{x} \equiv g(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample variance: $s^2 \equiv h(\mathbf{x}) \equiv \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$
- z-statistic of the sample mean: $z \equiv z(\mathbf{x}) = \frac{\overline{x}-\mu}{\mathbf{se}_{\overline{x}}}$
- t-statistic of the sample mean: $t \equiv t(\mathbf{x}) = \frac{\overline{x}-\mu}{\hat{\mathbf{se}}_{\overline{x}}}$

You've already seen the function g for computing the sample mean, which is identical to the **Mean** operation we learned about in the

chapter on descriptive statistics $g(\mathbf{x}) \equiv \overline{x} \equiv \mathbf{Mean}(\mathbf{x})$. When we calculate the variance s^2 of *sample* data instead of just any old data, the formula we use is a little different. In descriptive statistics, we used the formula ($\mathbf{Var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$), but in inferential statistics, we change the n to $n-1$. We'll explain why the formula is different in Section ???. For now, just remember that when calculating either the variance or the standard deviation using sample data, use the formula with $n-1$ instead of n. The z-statistic and the t-statistic are functions computed from the sample mean $\overline{x}$ with the purpose of "normalizing" the values obtained from different distributions and comparing them on the same scale. The values of these statistics are used as part of the z-test and the t-test, which are two common hypothesis testing procedures. The quantity $\mathbf{se}_{\overline{x}} \equiv \frac{\sigma}{\sqrt{n}}$ is called the *standard error* of the estimator $\overline{x} \equiv g(\mathbf{x})$. The quantity $\widehat{\mathbf{se}}_{\overline{x}} \equiv \frac{s}{\sqrt{n}}$ is an approximation for the standard error that is computed using the sample variance s^2 instead of the true population variance σ^2.

32.1.3 The theory behind estimation

Instead of a particular sample $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, let's now consider a *random* sample $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, where each X_i represents a random draw from the population. The value of the estimator computed from a random sample $f(\mathbf{X})$ is a random variable. Figure 32.2 illustrates the diagram for the theoretical analysis of random samples.

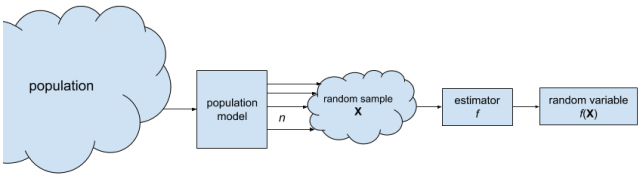

Figure 32.2: Using estimators in the context of statistical analysis of random samples. We start by building a probabilistic model for the distribution of values in the population. Using this model, we can perform the theoretical analysis of random samples $\mathbf{X}$ drawn from the population. In particular, we're interested in describing the variability of the random variable $f(\mathbf{X})$, which is output of an estimator computed on random samples $\mathbf{X}$. The probability distribution of this random variable $f(\mathbf{X})$ is called the *sampling distribution* of the estimator.

We use the following concepts as part of the hypothetical analysis of random samples from a population model.

- *Random sample of size n*: a theoretical concept that describes n independent, random draws from the population model: $\mathbf{X} \equiv \{X_1, X_2, \ldots, X_n\}$, where each $X_i \sim \mathbf{model}(\theta)$.
- *Estimator*: a function computed on random samples $\hat{\Theta} \equiv f(\mathbf{X})$. For example, the mean of a random sample: $\overline{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbf{Mean}(\mathbf{X})$. Note $\overline{X}$ is a random variable since it is computed from the random sample $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$. The probability distribution of the random variable $\overline{X}$ is called the *sampling distribution* of the sample mean.
- *Sampling distribution* of an estimator. The probability distribution that describes the random variable $\hat{\Theta}$.
- *Standard error* ($\mathbf{se}_{\hat{\theta}}$) of the estimator is the standard deviation of the sampling distribution for the estimator $\hat{\Theta}$. It's formula is $\mathbf{se}_{\hat{\theta}} \equiv \frac{\sigma}{\sqrt{n}}$. However, we almost never know the true population variance σ — to calculate it's value we would need to measure every member of the population. Instead we use the standard deviation of the sample s as an estimate for the standard deviation of the population σ. Replacing σ with s in the formula for standard error, we get $\hat{\mathbf{se}}_{\hat{\theta}} \equiv \frac{s}{\sqrt{n}}$ — an estimate of the standard error.

Let's look at the same estimators that we described in the previous section, but this time applied to *random* samples of size n:

- Random sample mean: $\overline{X} \equiv g(\mathbf{X}) \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$
- Random sample variance: $S^2 \equiv h(\mathbf{X}) \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$
- Sampling distribution of the z-statistic: $Z \equiv z(\mathbf{X}) = \frac{\overline{X}-\mu}{\mathbf{se}_{\overline{x}}}$
- Sampling distribution of the t-statistic: $T \equiv t(\mathbf{X}) = \frac{\overline{X}-\mu}{\hat{\mathbf{se}}_{\overline{x}}}$

The *sampling distribution* of an estimator describes the probability distribution of the estimator values computed on random samples of size n. By making assumptions about the probability distribution of the population and applying the basic laws of probability theory, we know the sample mean is going to be a normally distributed random variable $\overline{X} \equiv g(\mathbf{X}) = \mathcal{N}(\mu, \mathbf{se}_{\overline{x}}^2)$, where $\mathbf{se}_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$. We can similarly describe the variability of the sample variance $S^2 \equiv h(\mathbf{X})$. The sampling distributions of the z-statistic is the standard normal distribution $Z \sim \mathcal{N}(0,1)$. The sampling distribution of the t-statistic is described by Student's t-distribution which saw in Section 24.3.5.

Estimators are often used in conjunction with the expectation operator. The expected value the random sample mean $\overline{X} \equiv g(\mathbf{X})$ is equal to the population mean: $\mathbb{E}[\overline{X}] = \mu$, and the expected value

of the random sample variance $S^2 \equiv h(\mathbf{X})$ is equal to the population variance: $\mathbb{E}\left[S^2\right] = \sigma^2$. This where the name "estimators" comes from—the functions g and h are represent the computations we use to compute estimates for the population parameters μ and σ^2.

32.2 Formulas

32.2.1 Sample mean estimator

Given the sample $\mathbf{x} \equiv \{x_1, x_2, \ldots, x_n\}$, the sample mean is computed using the formula:

$$\overline{x} \equiv g(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \textbf{Mean}(\mathbf{x}).$$

You've seen this formula several times earlier in the book. The only new thing here is the interpretation—in this chapter we are using the quantity $\overline{x}$ to make an inference about the population mean μ.

32.2.2 Sample variance estimator

Sample variance is computed using the formula:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{n-1}\textbf{Var}(\mathbf{x}).$$

The quantity s^2 serves as an estimate of the population variance σ^2. The square root of the sample variance is called the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2}$.

Remember, the formula for the sample variance is different from the variance formula $\textbf{Var}(\mathbf{x})$ that we saw in Chapter 13. The denominator in the formula is $n-1$ and not n. We'll learn the reason for this in Section **??**.

32.2.3 Sampling distribution of the sample mean

The mean of a random sample $\mathbf{X} \equiv \{X_1, X_2, \ldots, X_n\}$ is defined as

$$\overline{X} \equiv g(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The probability distribution that describes the random variable $\overline{X}$ is called the *sampling distribution* of the estimator $\overline{x}$. The sampling distribution of the sample mean is normally distributed with mean $\mu_{\overline{x}}$

and variance $\mathbf{se}_{\overline{x}}^2$: $\overline{X} \sim \mathcal{N}(\mu, \mathbf{se}_{\overline{x}}^2)$. The law of large numbers tells us mean (expected value) of the random variable $\overline{X}$ is equal to the population mean μ. The standard deviation of $\overline{X}$ is called the *standard error* of the estimator $\overline{x}$ and computed using the formula

$$\mathbf{se}_{\overline{x}} = \frac{\sigma}{\sqrt{n}}.$$

Note the formula for the standard error depends on the population standard deviation σ and the square root of the sample size $\sqrt{n}$. The notation "**se**" stands for *standard error* and the subscript indicates the estimator whose standard error we're computing. For example, the standard error of the estimator $\hat{\theta}$ is denoted $\mathbf{se}_{\hat{\theta}}$.

Remember, the formula for the standard error $\mathbf{se}_{\overline{x}}$ assumes that the population standard deviation σ is known, which is not a very common case—usually the population standard deviation is an unknown quantity. We can compute an estimate for the standard error based on the sample standard deviation s instead of the population standard deviation σ:

$$\hat{\mathbf{se}}_{\overline{x}} = \frac{s}{\sqrt{n}}.$$

This is the quantity that we'll use most often in calculations going forward. The hat indicates the quantity $\hat{\mathbf{se}}_{\overline{x}}$ is estimated from a particular sample.

It's important to distinguish between standard error quantities $\hat{\mathbf{se}}_{\overline{x}}$ and $\mathbf{se}_{\overline{x}}$ from the population standard deviation σ and the sample standard deviation s. The population standard deviation σ measures the variability of the population, while the sample standard deviation s estimates σ by measuring the variability of the sample. The $\hat{\mathbf{se}}_{\overline{x}}$ and $\mathbf{se}_{\overline{x}}$ measure the variability of the estimator $\overline{X}$. The quantities are related (see formulas for $\mathbf{se}_{\overline{x}}$ and $\hat{\mathbf{se}}_{\overline{x}}$ above), and they are all a measure of variability, but they measure the variability of different things. The sample standard deviation s measures the variability *within* a sample. The $\hat{\mathbf{se}}_{\overline{x}}$ measures variability *between* samples.

32.2.4 Test statistics

Consider the population parameter θ and the estimator $\hat{\Theta} \equiv f(\mathbf{X})$ that has standard error $\mathbf{se}_{\hat{\theta}}$. Given a value of the estimator computed from a particular sample $\hat{\theta} = f(\mathbf{x})$, we calculate z-statistic using the following formula:

$$z \equiv z(\mathbf{x}) = \frac{\hat{\theta} - \theta}{\mathbf{se}_{\hat{\theta}}}.$$

The z-statistic measures the difference between the estimator value $\hat{\theta}$ computed from a particular sample and the true population pa-

rameter θ. Intuitively, you can think of the units of the z-statistic (sometimes called the z-score) as providing the "how many standard errors away from the mean" information. The z-statistic can be computed for all estimators whose sampling distribution is normally distributed.

The t-statistic is defined by the formula

$$t \equiv t(\mathbf{x}) = \frac{\hat{\theta} - \theta}{\widehat{\mathbf{se}}_{\hat{\theta}}},$$

which is similar to the formula for the z-statistic, but uses the *estimated* standard error $\widehat{\mathbf{se}}_{\hat{\theta}}$ in the denominator instead of $\mathbf{se}_{\hat{\theta}}$. The t-statistic also describes "how many estimated standard errors away from the mean" information.

These two test statistics are used in many statistical testing procedures that require computing standardized values in "number of standard deviations from the mean" units. Suppose we compute the estimator value $\hat{\theta} = f(\mathbf{x})$ from the particular sample $\mathbf{x}$ obtained from a population with parameter θ. The values of the the z-statistic and the t-statistic tell us how likely or unlikely it is to observe the value $\hat{\theta}$. We'll discuss statistical testing procedures in Chapter 33.

32.2.5 Confidence intervals

The *confidence interval* $\text{CI}_{(1-\alpha)}$ is a range of numbers that may contain the true value of some parameter of interest. A confidence interval is usually computed by taking the estimate for the parameter of interest, plus or minus some margin of error:

$$\text{estimate} \pm \underbrace{(\text{constant})_{\alpha} \cdot (\text{standard error of the estimator})}_{\text{margin of error}},$$

where the constant describes "how many standard errors away from the mean" parameter that depends on the required confidence level α.

Confidence intervals are usually written using interval notation with upper and lower limits. For example, the $(1-\alpha)$-confidence interval for the population sample mean of a normal distribution with known variance σ is given by

$$\text{CI}_{(1-\alpha)} = \left[\overline{x} - |z_{\alpha/2}| \cdot \mathbf{se}_{\overline{x}},\ \overline{x} + |z_{\alpha/2}| \cdot \mathbf{se}_{\overline{x}}\right].$$

where $z_{\alpha/2}$ denotes the value of the inverse CDF of the normal distribution $F_Z^{-1}(\alpha/2)$.

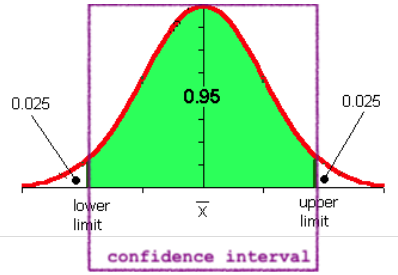

Figure 32.3: Illustration of a 95% confidence interval for the population mean computed from the estimate $\overline{x}$. We know the estimator $\overline{X}$ is normally distributed with mean μ and standard deviation $\mathbf{se}_{\overline{x}}$. Using these facts and the formula for the confidence interval provided above, we can find the region that contains 95% of the probability density function.

For cases where the population variance is not known, the confidence interval will be computed based on Student's T distribution and the estimated standard error:

$$\text{CI}_{(1-\alpha)} = \left[\overline{x} - |t_{\alpha/2,\nu}| \cdot \widehat{\mathbf{se}}_{\overline{x}}\,,\ \ \overline{x} + |t_{\alpha/2,\nu}| \cdot \widehat{\mathbf{se}}_{\overline{x}}\right].$$

The constant $t_{\alpha/2,\nu}$ denotes the value of the inverse CDF of the T distribution with ν degrees of freedom, $t_{\alpha/2,\nu} \equiv F^{-1}_{T_\nu}(\alpha/2)$. Note the structure of the confidence interval is the same ($\overline{x}$ plus or minus some multiple of the standard error), but we use the distribution T_ν instead of the normal distribution. Student's T distribution is similar in shape to the normal distribution but includes a "correction factor" to account for the fact we're using an estimate of the standard error $\widehat{\mathbf{se}}_{\overline{x}}$ instead of the true value $\mathbf{se}_{\overline{x}}$. We'll define the T distribution and explain the concept of degrees of freedom in Section ??.

32.2.6 Example

Consider the population of values that are normally distributed with mean $\mu = 70$ and standard deviation $\sigma = 10$. The probability model for random samples from the population is $X_i \sim \textbf{model}(\boldsymbol{\theta})$, where **model** is the normal distribution, and the model parameters are $\boldsymbol{\theta} = (\mu, \sigma^2) = (70, 100)$.

Let's place ourselves in the shoes of a scientists with a very limited budget who wants to estimate the population parameters by taking a *sample* of three values from the population. How much information about the population mean μ and standard deviation σ can the scientist learn from a sample of size $n = 3$? We'll investigate this question first theoretically by studying random samples, then show

the practical steps that the scientist could perform on a specific sample.

Random sample analysis We denote the a random sample of size $n = 3$ as follows $\mathbf{X} \equiv \{X_1, X_2, X_3\}$, where each X_i is randomly selected from the population model. Since the population is normally distributed with mean $\mu = 70$ and variance $\sigma^2 = 100$, the random values X_i have the distribution $X_i \sim \mathcal{N}(70, 100)$.

The sample mean estimator computed from a random sample of size $n = 3$ is

$$\overline{X} \equiv g(\mathbf{X}) = \tfrac{1}{3}\left(X_1 + X_2 + X_3\right),$$

and the sample variance estimator for the random sample is

$$S^2 \equiv h(\mathbf{X}) = \tfrac{1}{2}\Big[\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \left(X_3 - \overline{X}\right)^2\Big].$$

The random variable $\overline{X}$ is an estimate of the population mean $\mu = 70$. Similarly, the value of S^2 is an estimator for the population variance $\sigma^2 = 100$.

Note $\overline{X}$ is a random variable because it is computed from the random values X_1, X_2, and X_3. The central limit theorem tells us the distribution of the random variable $\overline{X}$ is a normal distribution with mean equal to the population mean μ and variance equal to $\frac{1}{3}$ as large as the population variance:

$$\overline{X} \sim \mathcal{N}\left(\mu, \tfrac{\sigma^2}{3}\right).$$

The random variable $\overline{X}$ describes the variability of the sample-mean estimates $\bar{x}$ computed from different samples of size $n = 3$ taken from the population. Using the formula for the *standard error* of this estimator (the standard deviation of the sampling distribution), we find $\mathbf{se}_{\bar{x}} = 10/\sqrt{3} = 5.77$.

Note the calculations apply to all possible samples taken from the population. We were able to predict the variability of the sample mean estimates is going to be $\mathbf{se}_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 5.77$ before collecting any samples!

Particular sample analysis Now suppose the scientist obtains a particular sample $\mathbf{x} \equiv \{64, 82, 73\}$ from the population. Let's look at some estimator values she can compute from this sample. We've intentionally chosen a very small sample size $n = 3$ to make the calculations easy to follow.

The first step for the scientist is to to computes the sample mean

$$\bar{x} \equiv g(\mathbf{x}) = \tfrac{1}{3}\left(64 + 82 + 73\right) = 73,$$

and the sample variance $s^2 \equiv h(\mathbf{X}) = 81$. The *estimator values* $\overline{x} = g(\mathbf{x})$ (sample mean) and $s^2 = h(\mathbf{x})$ (sample variance) can be combined to form the parameters estimate $\hat{\boldsymbol{\theta}} = (\overline{x}, s^2) = (73, 81)$, which describe the scientist's "best guess" about the true population parameters. This is the whole point of estimators—they are functions computed from the sample that allow us to find estimates for the population parameters.

But wait, there is more! Since we know the sampling distribution of the sample mean estimator $\overline{X} \equiv g(\mathbf{X})$, we can also report an estimate of our uncertainty about the value of $\overline{x} = 73$ we computed from the sample. The standard error for the sample mean estimator is $\mathbf{se}_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, which describes the variability of the estimates we are likely to obtain when computing samples of size $n = 3$. The population variance σ^2 is an unknown quantity, so we have to use the sample variance s^2 as an approximation for σ^2. Using the formula for the approximate standard error, we find $\hat{\mathbf{se}}_{\overline{x}} = \frac{s}{\sqrt{n}} = \sqrt{81}/\sqrt{3} = 5.12$. The standard error $\hat{\mathbf{se}}_{\overline{x}}$ is useful to compute so that we can quantify the uncertainty in our estimate for $\overline{x}$. In this case, our estimate for the population mean is 73 and the standard error of the sample mean is 5.12. This is much better: The standard error is an indication of the extent to which an estimate might vary with different samples. For that reason, when you report an estimate, you should always include its corresponding standard error.

The best way to express the uncertainty in our estimate of the population mean is to provide a confidence interval for our estimate. We can can compute a $(1-\alpha)$%-confidence interval for the population mean using the formula $\text{CI}_{(1-\alpha)} = [\overline{x} - |t_{\alpha/2,\nu=2}|\hat{\mathbf{se}}_{\overline{x}}, \overline{x} + t_{\alpha/2,\nu=2}\hat{\mathbf{se}}_{\overline{x}}]$, where $t_{\alpha/2,\nu=2}$ denotes the value of the inverse CDF of the T distribution with $\nu = n - 1 = 2$ degrees of freedom $F^{-1}_{T_{\nu=2}}(\alpha/2)$. Note we're using the values of the T distribution to compute the confidence interval because we're using the estimate standard error $\hat{\mathbf{se}}_{\overline{x}}$ (computed from s^2) and not the true standard error $\mathbf{se}_{\overline{x}}$ (computed from σ^2).

To find the 95%-confidence interval we choose $\alpha = 1 - 0.95 = 0.05$, and compute the value of the constant $t_{0.025,2} = -4.3$ (meaning $F_{T_2}(-4.3) = 0.025$). When computing confidence intervals, we're only interested in the "how for from the mean" information, so we take the absolute value of the t-statistic $|t_{0.025,2}| = 4.3$. In words, this number tells us that the true population mean can be anywhere up to 4.3 standard errors smaller or larger than our estimate 73. The

95%-confidence for the population mean is

$$\begin{aligned}\text{CI}_{0.95} &= [\overline{x} - 4.3\widehat{\mathbf{se}}_{\overline{x}},\ \overline{x} + 4.3\widehat{\mathbf{se}}_{\overline{x}}] \\ &= [73 - 4.3 \cdot 5.12,\ 73 + 4.3 \cdot 5.12] \\ &= [50.64,\ 95.36].\end{aligned}$$

Computing the confidence interval for our estimate provides useful information about our uncertainty of the population mean. The very wide confidence interval tells us not to trust our estimate for the population mean $\overline{x} = 73$ too much—the true population mean could be anywhere in the range $[50.64, 95.36]$, which is very broad. That's what you get when you try to do statistics with very small sample sizes.

32.3 Understanding why estimators work

Many of the formulas and probability calculations that we use in statistics are computed based on estimates and the analysis of the estimator's sampling distribution. It's easy to grasp the notion of the estimator value $\hat{\theta} = f(\mathbf{x})$ computed from a particular sample $\mathbf{x}$. It's much more difficult to think about and the sampling distribution of estimator $\hat{\theta}$, which is a theoretical construct that describes the random variable $\hat{\Theta} = f(\mathbf{X})$ computed from random samples $\mathbf{X}$ taken from the population. What does the sampling distribution of an estimator describe exactly? What does it look like? As with all things mathematical, it's always useful to look at a picture before looking at the math formulas.

Figure 32.5 shows multiple repetitions of the random sampling procedure for three different sample sizes. The first row corresponds to samples of size $n = 3$ taken from the population. The values of the sample mean estimator $\overline{x} = g(\mathbf{x})$ computed from each sample oscillate around the centreline which represents the true population mean μ. In the second row we see the same situation repeated with samples of size $n = 15$, and the third row shows samples of size $n = 200$. Note how the variability of the sample mean estimates decreases when we take larger samples.

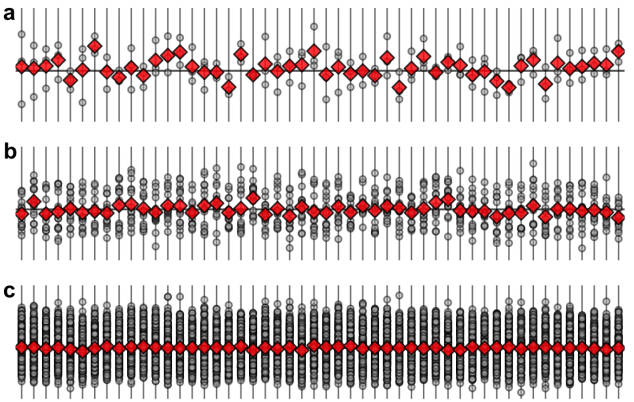

Figure 32.4: The data from multiple independent random samples taken from the population are shown horizontally. The grey circles in each sample correspond to the data points $\{x_{k1}, x_{k2}, \ldots, x_{kN}\} = \mathbf{x}_k$, while the hollow diamond shapes indicate the sample mean $\overline{x} = \frac{1}{n}\sum_i x_i$ within each sample. Three cases are shown: **a)** sample size $n = 3$, **b)** sample size $n = 15$, and **c)** $n = 200$. What is the trend you see about the diamond shapes when you go from **a)**, to **b)**, and **c)**?

The sample mean values computed in each sample in the above figure (diamond shapes) are instances of the random variable $\overline{X}$. The sampling distribution is an abstract theoretical construct that describes the variability of the estimator $\overline{X}$. We can visualize the sampling distribution by plotting a histogram for the different values of $\overline{x}$ obtained from repeated sampling, as illustrated in Figure 32.5.

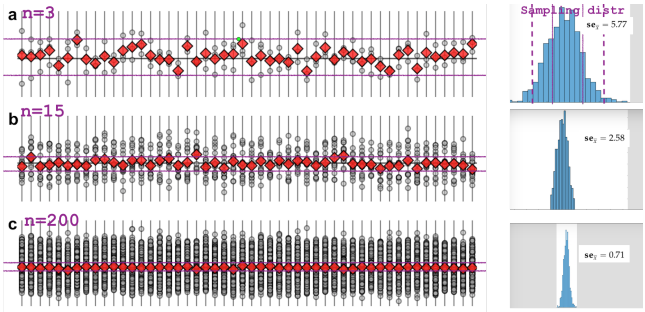

Figure 32.5: Visualization of the sampling distribution for the sampling distribution of the sample mean $\overline{x}$ computed from 50 different random samples of size $n = 3$, $n = 15$, and $n = 200$. The solid indicate the interval $[\overline{x} - \mathbf{se}_{\overline{x}}, \overline{x} + \mathbf{se}_{\overline{x}}]$, where $\mathbf{se}_{\overline{x}}$ is the standard error of the estimator $\overline{x}$.

Note the shape of the histogram always looks roughly like a normal

distribution, and the larger the samples get the narrow the distribution of estimates becomes. This approximately-normal behaviour is predicted by the central limit theorem, which states that the sampling distribution of any estimator will "approach" the normal distribution for large enough samples, In other words, the central limit theorem tells us

$$\overline{X} \sim \mathcal{N}\left(\mu, \mathbf{se}_{\overline{x}}^2\right),$$

where $\mathbf{se}_{\overline{x}}$ is the standard error of sample mean estimator.

The standard error $\mathbf{se}_{\overline{x}}$ is a useful way to quantify the variability of the estimator values. Looking at the right side of Figure 32.5 we can observe that the standard error of the sample mean estimator for samples of size $n = 3$ is $\mathbf{se}_{\overline{x}} = 5.77$. When using samples of size $n = 15$ the standard error decreases to $\mathbf{se}_{\overline{x}} = 2.58$, and using samples of size $n = 200$ the standard error is $\mathbf{se}_{\overline{x}} = 0.71$.

The central limit theorem provides us with a formula for standard error of the sampling distribution:

$$\mathbf{se}_{\overline{x}} = \frac{\sigma}{\sqrt{n}},$$

where σ is the population standard deviation. This means the standard error of our estimates decreases when we collect larger samples. Specifically, the standard error decreases as $\frac{1}{\sqrt{n}}$ as n increases.

Note this behaviour applies to all probability distributions: the standard error of any estimator decreases as the sample size gets larger, regardless of the shape of the population's distribution. This is called the *asymptotic normality* property of estimators. We'll talk more about that in Section 32.4.5.

32.4 Estimator properties

In this section we'll introduce some new terminology for describing the properties of estimators.

Consider the estimator $f : \mathcal{X}^n \to \mathbb{R}$ that is used to compute estimator values $\hat{\theta} \equiv f(\mathbf{x})$. The sampling distribution of this estimator is defined as a random variable obtained by applying the function f to a random sample $\mathbf{X}$:

$$\hat{\Theta} = f(\mathbf{X}).$$

The random variable $\hat{\Theta}$ describes the variability of the different estimator values $\hat{\theta}$ that we are likely to observe if we were to repeatedly draw new samples of size n from the population. The probability distribution that describes $\hat{\Theta}$ is called "the sampling distribution of the estimator $\hat{\theta}$." In it customary in statistics to refer to the random

variable $\hat{\Theta}$ as "the estimator $\hat{\Theta}$" instead of "the values of the estimator f computed from random sample $\mathbf{X}$," which would be the more precise terminology.

32.4.1 Estimator bias

An estimator is *unbiased* if the mean of its sampling distribution is equal to the value of the population parameter being estimated. The estimator $\hat{\Theta}$ is an *unbiased estimator* if $\mathbb{E}[\hat{\Theta}] = \theta$. If the estimator is not unbiased then the difference $\mathbb{E}[\hat{\Theta}] - \theta$ is called the *bias*.

Suppose that X is a random variable with mean μ and variance σ^2. Taking random samples $X_1, X_2, \ldots, X_n$ from the population, we will now show that the sample mean $\overline{X}$ is an unbiased estimators of μ:

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n}{n}\mu = \mu.$$

The expected value of each random variable X_i is μ, hence the sum of n independent copies of X_i divided by n equals the population mean μ. We can therefore say "$\overline{X}$ is an unbiased estimator for the population mean μ."

We can perform a similar calculation for the sample variance, and after some calculations we obtain:

$$\mathbb{E}[S^2] = \tfrac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \sigma^2,$$

which tells us S^2 is an unbiased estimator of the population variance σ^2.

Note the formula for the sample variance estimator is $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$ with normalization factor $\frac{1}{n-1}$ in front of the summation instead of $\frac{1}{n}$. As you can see from the above calculation, this is not some arbitrary choice we make but a consequence of probability calculations. A sample variance formula with normalization factor $\frac{1}{n}$ would systematically underestimate the population variance.

32.4.2 Estimator variance

The variance of an estimator $\hat{\Theta}$ for the population parameter θ is defined by the following formula

$$\mathbb{V}[\hat{\Theta}] \equiv \mathbb{E}\Big[\big(\hat{\Theta} - \mathbb{E}[\hat{\Theta}]\big)^2\Big].$$

This formula computes the expectation of the squared difference between the estimator $\hat{\Theta}$ and its expected value $\mathbb{E}[\hat{\Theta}]$. You will recognize this expression as an instance of the usual variance formula applied to the random variable $\hat{\Theta}$. The meaning of this quantity describes the the variability of the random estimates $\hat{\theta} = f(\mathbf{x})$ that we are likely to compute from different samples of size n from the population.

Two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ have distributions centred in the same spot, but could still have a different variance. Given the choice between the two estimators, we will prefer the estimator with smaller variance.

TODO: example of two unbiased estimators with difference variance

32.4.3 Standard error of an estimator

The *standard error* of an estimator $\hat{\Theta}$ is the square root of the estimator variance $\mathbf{se}_{\hat{\theta}} = \sqrt{\mathbb{V}[\hat{\Theta}]}$. All self-respecting statisticians will report estimates of the uncertainty associated with each of value $\hat{\theta}$ they report. For example, if the statistician computes the estimate $\hat{\theta}$ for the model parameter θ, it is their obligation to also report an estimate of the standard error $\widehat{\mathbf{se}}_{\hat{\theta}}$.

In the formulas section above (Section 32.2), we showed two formulas for computing the standard error of the sample mean estimator $\overline{X}$ for samples of size n. If the population variance σ^2 is known, we can compute the exact standard error of the estimator using the formula $\mathbf{se}_{\hat{\theta}} = \frac{\sigma}{\sqrt{n}}$. If the population variance is unknown, we can instead compute the estimated standard error using the formula $\widehat{\mathbf{se}}_{\hat{\theta}} = \frac{s}{\sqrt{n}}$, where the sample standard deviation s is used as a substitute for the population standard deviation σ. The estimated standard error $\widehat{\mathbf{se}}_{\hat{\theta}}$ is used in calculating test statistics, confidence intervals, and other quantities used in statistical analysis procedures.

32.4.4 Mean squared error of an estimator

The *mean squared error* (MSE) of an estimator is a useful metric that combines the notions of bias ($\mathbb{E}[\hat{\Theta}] - \theta$) and variance ($\mathbb{V}[\hat{\Theta}]$). The mean squared error of an estimator $\hat{\Theta}$ is the expected value of the squared difference between $\hat{\Theta}$ and the population parameter θ:

$$\text{MSE}(\hat{\Theta}) = \mathbb{E}[(\hat{\Theta} - \theta)^2].$$

By rewriting this expression $(\hat{\Theta} - \theta)$ as $(\mathbb{E}[\hat{\Theta}] - \theta) + (\hat{\Theta} - \mathbb{E}[\hat{\Theta}])$ and doing some algebra calculations we can arrive at the following equiv-

alent formula:

$$\begin{aligned}\text{MSE}(\hat{\Theta}) &\equiv \mathbb{E}[(\hat{\Theta} - \theta)^2] \\ &= \underbrace{\Big(\mathbb{E}[\hat{\Theta}] - \theta\Big)^2}_{\text{bias term}} + \underbrace{\mathbb{E}\Big[\big(\hat{\Theta} - \mathbb{E}[\hat{\Theta}]\big)^2\Big]}_{\text{variance term}} \\ &= (\text{bias})^2 + \mathbb{V}[\hat{\Theta}].\end{aligned}$$

Note how mean squared error of the estimator decomposes into a bias term and a variance term. The bias term measure how much the estimator is "consistently off" from the true value θ on average. The variance term describes the variability of the estimates $\hat{\theta}$ we are likely to observe when selecting samples of size n from the population. Bias and variance are two central concepts of statistics: we generally prefer to use unbiased estimators and aim to quantify the variance of each estimator by computing its standard error.

In some cases, a biased estimator may be preferred over an unbiased one because its mean squared error is smaller. In particular, the "best" choice of estimator to use might depend on the sample size. For example it might be better to use a biased estimator with small variance rather an unbiased estimator with large variance. An estimator is said to be more *efficient* than another if it has smaller mean squared error. An estimator having mean squared error less than any other is called an *optimal estimator*.

TODO: example of biased estimator with smaller MSE overall

Decomposing the estimator "error" into a bias term and a variance term is a very useful way to characterize the "overall quality" of any probabilistic model. We'll see these concepts again in the machine learning chapters, where we'll talk about the *bias-variance tradeoff* that we must make when choosing a machine learning model for a given ML task.

32.4.5 Asymptotic normality

Provided the sample size is large enough, the sampling distribution of any estimator that computes the sum of independent random variables will be a normal distribution. We have used this result several times in this chapter already, but it's worth going over it one more time.

Consider first a normally distributed population $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ with mean μ_x and variance σ_x^2, and a random sample $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ that consists of n independent draws from the population model.

The sample mean estimator $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is normally distributed:

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{n}\right),$$

with mean equal to the population mean and variance equal to $\frac{\sigma_x^2}{n}$. Note this is an exact formula that can be derived from the properties of a sum of normally distributed random variables and not an approximation. The formula is true for all n, including the case $n = 1$. The standard error of the estimator $\overline{X}$ decreases as the sample size n get larger: $\mathbf{se}_{\overline{x}} \equiv \mathbb{V}[\overline{X}] = \frac{\sigma}{\sqrt{n}}$. This is the behaviour we observed in Figure 32.5 (see page 13).

Now consider another population model $Y \sim \mathbf{model}(\theta)$, where **model** is some probability distribution (not necessarily normal) and θ describes the model parameters. The population mean for this model is defined as $\mu_y = \mathbb{E}[Y]$ and the population variance is $\sigma_y^2 = \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big]$. The central limit theorem (Theorem **??** on page **??**) tell us that, under some general conditions about the model distribution **model**, the sample mean estimator $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ will be approximately normally distributed for large enough n:

$$\lim_{n\to\infty} \overline{Y} \sim \mathcal{N}\left(\mu_y, \frac{\sigma_y^2}{n}\right).$$

The above formula is telling us three important things we need to know about the sampling distribution of the estimator $\overline{Y}$. First it tells us the sampling distribution is basically a normal distribution, which simplifies all kinds of probability calculations that we might want to perform as part of a statistical analysis, since normal distributions are some of the easiest models to work with. The population distribution **model** could be any complicated function, but we only need to do probability calculations using the normal distribution to perform the statistical analysis procedures that involves sample mean $\overline{y}$ estimates. Second it tells us the expected value of the random variable $\overline{Y}$ is equal to the population mean μ_y, which is to be expected since we want to use $\overline{Y}$ as an estimator for the population mean. The third piece of information that the central limit theorem provides us with, is the formula for that the variance of the estimator $\overline{Y}$, which we usually express as a formula for the standard error of the estimator $\mathbf{se}_{\overline{x}} \equiv \frac{\sigma_y}{\sqrt{n}}$. The sampling distribution of the sample mean $\overline{Y}$ from *any* distribution is normally distributed with standard error decreasing as the sample size increases.

The notion of "large enough" sample size depends on the particular statistical analysis procedure you want to compute. This is why

all statistical analysis procedures include an "assumptions checklist" that clearly states the sample size required to use the procedure for a given type of population model. In Chapter 33 we'll talk more about how to choose appropriate sample size n for a given experiment. For now, all you can need to remember is that when n is large enough (think $n \geqslant 30$), we can treat the "approximately normal" distribution of the estimator $\overline{Y}$ as it were exactly normal:

$$\overline{Y} \sim \mathcal{N}\left(\mu_y, \mathbf{se}_{\overline{x}}^2\right),$$

with standard error given by $\mathbf{se}_{\overline{x}} \equiv \frac{\sigma_y}{\sqrt{n}}$.

In summary, statistics is made possible thanks to the central limit theorem.

Exercises

32.5 Sampling distributions reference

In this section we'll discuss three of the most important probability distributions. You've already seen the mathematical definitions and computational procedures for computing the values of these functions in Section 24.3, but it's worth reviewing the the properties of these distributions using the new terminology of estimators (random variables) and estimator values (numbers).

For each of these distributions, you need to become fluent at computing values of their CDF and their inverse CDF functions. Assuming the distribution of the random variable X with probability density function f_X and cumulative density function F_X, the two most common type of calculations you'll need to perform are the following:

- Given a particular value of the random variable x, you need to know how to calculate the probability of the event $\{X \leqslant x\}$. This corresponds to the value of the cumulative density function $F_X(x) \equiv \Pr(\{X \leqslant x\}) = \int_{-\infty}^{x} f_X(x')\,dx'$.
- Alternatively, if you're given some probability value q, you need to be able to find the value of the inverse cumulative density function for that probability $x_q = F_X^{-1}(q)$. This corresponds to solving the equation $q = F_X(x_q) = \Pr(\{X \leqslant x_q\}) = \int_{-\infty}^{x_q} f_X(x')\,dx'$ for the unknown x_q.

The tables in Appendix YY contain values of inverse CDF F_X^{-1} for the standard normal, Student's t, and χ^2 distributions. You can

use these tables to lookup both the values $F_X^{-1}(q)$ and $F_X(x)$ in cases when you don't have access to a computer, for example when solving exercises and problems in the end of this chapter using pen and paper, sitting in a park with a coffee and your phone switched into airplane mode to avoid distractions. The lookup-values-in-a-table approach is also the one you're expected to use on statistics exams.

We'll also provide the names of the Excel, Python, and R functions you can use to obtain these values of $F_X^{-1}(q)$ and $F_X(x)$ in cases when you have access to a computer. It's up to you to choose you favourite method for computing probabilities, but it's a good idea to get *really* comfortable with at least one of these three options, because computing values of the standard normal, Student's t, and χ^2 distributions will be used a lot in the upcoming statistics chapters.

32.5.1 The standard normal distribution

The sampling distribution of any estimator $\hat{\theta}$ that computes the sum of independent random variables is a normal distribution. Since the formula for the sample mean estimator $\overline{x}$ fits this criterion, we'll see the normal distribution come up again and again every time we compute the sampling distribution of the sample mean.

Recall that any normal random variable X can be transformed to the *standard normal* distribution $Z \sim \mathcal{N}(0,1)$ using the transformation $z = \frac{x-\mu_x}{\sigma_x}$. This means any probability computation that we might want to perform on the random variable X, there is an equivalent calculation we can perform using the transformed value Z. This all-normal-distributions-are-the-same property is a great computation simplification since we only need to know how to compute values of one normal distribution—the standard normal.

Consider now a normally distributed estimator $\hat{\Theta} \sim \mathcal{N}(\mu_\theta, \mathbf{se}^2_{\hat{\theta}})$, with mean μ_θ and standard error $\mathbf{se}_{\hat{\theta}}$ The value of the z-statistic (z-score) for a given estimator value $\hat{\theta} = f(\mathbf{x})$ is obtained using the transformation

$$z = \frac{\hat{\theta} - \theta}{\mathbf{se}_{\hat{\theta}}},$$

where θ is the population value θ.

Probability calculations TODO: explain given z calculate p

TODO EXAMPLE: CDF calculation

TODO: explain given q calculate z_q

You can compute it by looking up the value in Table XX, or using the formula z_q `=NORM.INV(q,0,1)` in spreadsheet software, or calling z_q `=??` in R or z_q `=norm.ppf(q,0,1)` in Python.

The z statistic is used whenever the sampling distribution of the estimator $\hat{\theta}$ is normally distributed with standard error $\mathbf{se}_{\hat{\theta}}$. The normal distribution is used in conjunction with the values of the z-statistic in order to perform statistical analysis for all normally distributed estimators.

TODO EXAMPLE: inverse CDF calculation (confidence interval)

32.5.2 Student's t distribution

When the population variance σ^2 is an unknown quantity, we use the sample variance s^2 as an estimate for it.

TODO: mention it's a family of distr. and define df

$$t = \frac{\hat{\theta} - \theta}{\hat{\mathbf{se}}_{\hat{\theta}}},$$

Student's t distribution is used in conjunction with the t-statistic to perform statistical analysis in case where the population variance is estimated from a sample.

The sampling distribution of the estimator is normally distributed, but we don't know the true value of the standard error $\mathbf{se}_{\overline{x}} = \frac{\sigma^2}{\sqrt{n}}$ and instead use the estimated standard error $\hat{\mathbf{se}}_{\overline{x}} = \frac{s^2}{\sqrt{n}}$.

Student's t-distribution has an interesting history...
TODO: say it comes from "industry"
JOKE: the employer (Guinness brewing company) did not let him use his name for fear that the news of a "bad batch" of Guiness that is thrown out — Irish ppl would be outraged :) and storm the factory being like "give me the bad batches...."
TODO: mention general concept of Studentization `https://en.wikipedia.org/wiki/Studentization`.

TODO EXAMPLE: calculate the probability of observing the test statistic $t = 5$ for the sample mean $\overline{x}$ computed from samples of size $n = 9$ taken from a population with mean μ and variance σ^2.

TODO EXAMPLE inv: remind readers of the $CI_{1-\alpha}$ calculation in the Example

32.5.3 The χ^2 distribution

The sampling distribution of any estimator that computes the sum of squares independent random variables is a χ^2 distribution. The superscript 2 gives us a hint that the quantity has something to do with squares. The Greek letter χ is spelled "chi" (rhymes with "bye"), so "χ^2" is read "chi squared."

There is a whole family of χ^2-distributions defined by the different values of the parameter ν, which represents the *number of degrees of freedom* of the distribution. The degrees of freedom parameter is sometimes denoted "df." We'll use the symbol "ν" in math equations and "`df`" in code examples.

Figure 24.4 (page 107) shows several chi-square distributions for different values of the degrees of freedom parameter ν.

Used in conjunction with the sample variance estimator S^2

TODO EXAMPLE: calculate the probability of variance ... greater than assuming ...

32.5.4 More estimators

In addition to the essential, must-know estimator formulas we gave in Section 32.2 above, there are a number of "secondary" estimator formulas you should be aware of, as they will come up in certain calculations.

32.5.5 Functions of other estimator

Since estimators are random variables, all the algebra rules that apply for random variables also apply for estimators. For example, we can combine two existing estimators $\hat{\Theta}_1 \sim \mathcal{N}(\mu_1, \mathbf{se}^2_{\hat{\theta}_1})$ and $\hat{\Theta}_2 \sim \mathcal{N}(\mu_2, \mathbf{se}^2_{\hat{\theta}_2})$ to form a new estimator that computes difference of their values

$$\hat{D} \equiv \hat{\Theta}_1 - \hat{\Theta}_2.$$

Using the general rule of probability for sums and differences of normally distributed random variables, we know the sampling distribution of the estimator $\hat{D}$ is a normal distribution: $\hat{D} \sim \mathcal{N}(\mu_1 - \mu_2, \mathbf{se}_{\hat{d}})$. The expected value of $\hat{D}$ is equal to the difference between the population means $\mathbb{E}[\hat{D}] = \mu_1 - \mu_2$, and standard error of this estimator is

$$\mathbf{se}_{\hat{d}} = \sqrt{\mathbf{se}^2_{\hat{\theta}_1} + \mathbf{se}^2_{\hat{\theta}_2}}.$$

The difference-between-two-quantities estimator $\hat{D}$ is used in many statistical analysis procedures where we want to compare two groups.

32.5.6 Difference of means estimator

Consider a sample $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ of size n taken from one group or population, and the sample $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ of size m that is taken from a different group or population. The value of the difference of sample means estimator computed from these two samples

is

$$\hat{d} \equiv \overline{x} - \overline{y},$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means computed from each group. The estimator value $\hat{d}$ is an estimate for the difference between the group means μ_x and μ_y. The sampling distribution of the estimator $\hat{D}$ is

$$\hat{D} \sim \mathcal{N}(\mu_x - \mu_y, \mathbf{se}_{\hat{d}}^2),$$

where standard error is given by the formula

$$\mathbf{se}_{\hat{d}} = \sqrt{\mathbf{se}_{\overline{x}}^2 + \mathbf{se}_{\overline{y}}^2} = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}.$$

In cases when the populations variances are known and equal $\sigma_x = \sigma_y = \sigma$, the standard error formula simplifies to $\mathbf{se}_{\hat{d}} = \sigma\sqrt{1/n + 1/m}$.

Computing the true value of the standard error $\mathbf{se}_{\hat{d}}$ requires knowledge of the population variances σ_x^2 and σ_y^2, which is rarely the case. More often we'll use the values of the sample variances s_x^2 and s_y^2 to obtain the estimated standard error:

$$\hat{\mathbf{se}}_{\hat{d}} = \sqrt{\hat{\mathbf{se}}_{\overline{x}}^2 + \hat{\mathbf{se}}_{\overline{y}}^2} = \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}.$$

Note the formula for $\hat{\mathbf{se}}_{\hat{d}}$ is the same as the formula for $\mathbf{se}_{\hat{d}}$, but we have "plugged in" s_x instead of σ_x and s_y instead of σ_y. This is an instance of the general "plug in principle," which is used a lot in statistics.

32.5.7 Proportion estimator

Suppose we're studying a population variable with two possible values like 1/0, YES/NO, hard disk works/hard disk failed, etc. We could model the population using $X \sim \texttt{Bernoulli}(p)$ where p the population parameter that describes the proportion of individuals that have the characteristic of interest.

Given the sample $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ of size n taken from this population, the proportion estimator is defined as

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \mathbf{Count}(\mathbf{x}),$$

where $\sum x_i$ represents the count of individuals within the sample that have the characteristic. The estimator value $\hat{p}$ is an approximation for the true population parameter p.

Applying the central limit theorem to the random variable $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (proportion estimator for a random sample $\mathbf{X}$), tells us the sampling distribution of the proportion is approximately normally distributed with mean p and variance $p(1-p)$. In other words,

$$\hat{P} \sim \mathcal{N}(p, p(1-p)),$$

which tells us the proportion estimator $\hat{P}$ is approximately normally distributed with mean equal to the population parameter p and standard error given by

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

The estimated standard error is given by

$$\hat{\mathbf{se}}_{\hat{p}} = \frac{s}{\sqrt{n}},$$

where $s^2 = \hat{p}(1-\hat{p})$ is the sample variance.

Normal approximation valid when $np \geqslant 10$ and $n(1-p) \geqslant 10$ (back reference to prob theory: Discussion where show normal approx to binomial)

We'll use this result to when making inferences about population proportion parameters.

TODO: CI for proportion estimator

32.5.8 Sampling distribution of the sample variance

If S^2 is the variance of a random sample of size n from a normal population having variance σ^2, then the sampling distribution of $\frac{(n-1)S^2}{\sigma^2}$ is χ^2 with $n-1$ degrees of freedom.

We use this result for inferences concerning the population standard deviation σ.

TODO EXAMPLE: repeat example calculation of S^2 example above? new one? cut?

32.5.9 Linear correlation

The population quantity ρ measures... The estimator $\hat{\rho}$ (sometimes denoted r)

$$\hat{\rho} \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

is an estimator for the population linear correlation $\rho = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i-\mu_X}{\sigma_X}\right)\left(\frac{Y_i-\mu_Y}{\sigma_Y}\right)$

32.6 Discussion

32.6.1 Random sampling

All the results and formulas presented in this chapter rely on fundamental assumption that the samples collected from the population are obtained through the process of random selection. Basically, *random selection* means you're using a procedure that makes sure every member of the population has an equal chance of being chosen. The random selection includes the following aspects:

- *Independent samples*: the value of value in the sample doesn't influence the value of another member.
- *Representative samples*: we're sampling uniformly from the entire distribution and not disproportionately preferring one subgroup
- *No selection bias*: all members of the population must have a chance.
- *Small measurement error*: the quantity you're measuring must be accurate. This is often a problem with questionnaires.

You must watch out for these assumptions when collecting data for real-world statistical experiments, because mathematical machinery used might no longer be valid if any the assumptions are violated. We'll talk more about experimental design and practical considerations in Chapter ??.

32.6.2 Applications

Many statistical inference procedures that will be discussed in the next chapter boil down to the calculation of an estimator value $\hat{\theta}$ from a particular sample $\mathbf{x}$, and measuring the likelihood of such value to occur by chance under a given sampling distribution, $\hat{\Theta} \sim \mathcal{N}\left(\mu_\theta, \mathbf{se}^2_{\hat{\theta}}\right)$.

We'll now describe some calculation based on estimators to illustrate the concepts involved, but reserve the detailed explanations for the next chapter. In all of these applications, estimators are the computational building block and the only difference is the interpretation we give to the different estimator values obtained.

Test statistics A test statistics is a particular type of estimator that is used as part of a hypothesis testing procedure. The general structure of test statistic is to *normalize* an estimator value by subtracting the expected population parameter and dividing but the standard error:

The z-statistic ($z \equiv \frac{\hat{\theta}-\mu_\theta}{\mathbf{se}_{\hat{\theta}}}$) and the t-statistic ($t = \frac{\hat{\theta}-\mu_\theta}{\widehat{\mathbf{se}}_{\hat{\theta}}}$) are two of the most commonly used test statistics.

p-values The p-value associated with the value of a test statistic is a probability calculation that tells us how likely it is to observe this value of the test statistic, under some assumptions about the population.

Suppose we've obtained the estimator value $\hat{\theta}$ computed from some sample $\mathbf{x}$. We can "normalize" the estimator value $\hat{\theta}$ by transforming it to the standard test statistic $z \equiv \frac{\hat{\theta}-\mu_\theta}{\mathbf{se}_{\hat{\theta}}}$. The p-value is the probability of observing a value of the test statistic equally or more extreme than the value z:

$$p = \int_z^\infty f_Z(x)\,dx = 1 - F_Z(z).$$

This p-value can be extrapolated by looking at Table ?? or computed using ...

The general idea behind the p values is to measure how likely or unlikely it is to observe a given value of the test statistic. If the observed value of the estimator $\hat{\theta}$ is very unlikely to occur by chance under the sampling distribution this will give us reason to doubt the initial hypothesis...

Critical values A *critical value* is a value of the test statistic that can be determined ahead of time in order to form a decision rule. For example, we can choose a rejection level α and find the corresponding $\text{CV}_{\hat{\theta}} = F^{-1}(\alpha)$. Later on after we collect data and compute the value of the estimator $\hat{\theta}$ we just have to compare it to the critical value $\text{CV}_{\hat{\theta}}$ to decide whether to reject a hypothesis.

Confidence intervals The confidence interval for any estimate is a range of numbers that includes plausible values for a parameter we're estimating. The confidence level 95% ($\alpha = 0.05$) gives the overall success rate of the method for calculating the confidence interval. The confidence interval for estimating a population parameter θ based on the estimate value $\hat{\theta}$ has the form

$$\text{CI}_{(1-\alpha)} = \left[\hat{\theta} - F^{-1}(\alpha/2)\cdot\mathbf{se}_{\hat{\theta}}\,,\;\; \hat{\theta} + F^{-1}(\alpha/2)\cdot\mathbf{se}_{\hat{\theta}}\right].$$

where F^{-1} is the inverse CDF of the sampling distribution for the estimator $\hat{\theta}$.

Effect size estimates An *effect size* estimate is an estimate for a particular real-world quantity that we're interested in. For example, consider an educational experiment that wants to compare the average grade between two groups of students. Suppose we collect samples of size n from the two groups, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$, and calculate the sample mean within each sample $\overline{x}$ an $\overline{y}$. The estimator $\hat{d} = \overline{x} - \overline{y}$ measures the difference of between sample means, and it serves as an estimate of the difference between population means μ_x and μ_y.

The value of the estimator $\hat{d}$ can be used to evaluate if there is a difference between the population means. If the observed value of $\hat{d}$ is significantly different from zero, we can claim that there is a difference between the groups. More importantly, the value of the estimate $\hat{d}$ can be reported as a measure of *how much* grades differ between the groups, which is much better than the basic statement "grades have improved." We can also compute a confidence interval of the effect size.

We'll talk more about p-values, critical values, confidence intervals, and effect size estimators in the next chapter when we'll learn about null hypothesis significance testing procedure.

* * *

Let's summarize what we learned in this chapter. We introduced the notion of estimator $\hat{\theta} \equiv f(\mathbf{x})$ computed on a sample $\mathbf{x}$, and its sampling distribution $\hat{\Theta} \equiv f(\mathbf{X})$, which is obtained through the statistical analysis of random sample from the population. In practice, all the estimator knowledge you'll need is to remember is the formula for the standard error of each estimator $\mathbf{se}_{\hat{\theta}}$ (or $\widehat{\mathbf{se}}_{\hat{\theta}}$) and know how to compute probabilities, test statistics, and confidence interval. Most of the statistical analysis techniques you'll perform in the coming chapters will require identifying the appropriate estimator to use in each situation and applying the appropriate formulas. This is the beauty of statistics—we do a lot of theoretical analysis up front in order to make the practical analysis as simple as plugging numbers into formulas.

TODO: Forward reference to NHST and PARAMETRIC chapters "stat test subtype" concept, where each subtype comes with pre-packaged formulas

32.7 Links

https://en.wikipedia.org/wiki/Bessel%27s_correction

32.8 Estimators problems

General feedback from Robyn in no particular order:

- General impression: good! All the content is here, but I think some parts are way more in depth than needed, and other parts need more detail. I've tried to indicate these in the rtodo comments.
- Do you think it would be possible to just talk about estimating population parameters without bringing up modelling (yet)? I think it would simplify things a lot.
- I think we should focus more on making sure that by the time the reader finishes this chapter, they are able to use sample to make a best guess about a population.
- Readers should be familiar with the normal distribution BEFORE this chapter.
- Maybe we'd like to introduce the "68–95–99.7 rule" here or before this chapter. I think this concept would help readers understand why standard error and the normal distributions are so useful.
- Somewhere here we should introduce how to plot uncertainty (e.g. confidence intervals)
- Suggest that we avoid using the general "statistics" or "statistical analysis" when we mean something more specific (e.g. "inference", "estimation"?).
- If we keep CLT and LLN in a section called "extra topics", then readers might skip it before they get to this chapter. Maybe we should review those two concepts more thoroughly here.