Statistical testing vs. interval estimation

Chp. 8 focused on estimating a parameter (μ, p, etc.) and placing some bounds on the error of estimation.

- Can describe the population more accurately - knowing the mean income of a population gives us an idea of the average person's economic position
- Can report on the status of a situation - the difference in the percentage of voters supporting Bush and those supporting Gore ($p_1 - p_2$) may be of interest just before an election?

Chp. 10 focuses on statistical testing.

- We have hypotheses about the population, and we use a random sample to help us decide which is accurate
- We may be testing a new medication/treatment, a new scientific theory, etc.
- Is the value of a parameter equal to a particular value, or is it larger? smaller? just different?

What are the hypotheses?

To begin, we need to formulate our two hypotheses.

Null hypothesis

- Expressed in the form $\theta = \theta_0$, where θ is a population parameter and θ_0 is a specific value hypothesized for that parameter
- Denoted H_0
- H_0 is the hypothesis to fall back on if we don't have enough evidence to support H_A

Alternative hypothesis

- Expressed in one of the three forms: $\theta \neq \theta_0$, $\theta > \theta_0$, $\theta < \theta_0$,
- Denoted H_A (or occasionally H_1)
- The "burden of proof" is on H_A

Types of errors

Since we have two hypotheses, there are two errors that we could make.

Type I error

- Occurs when H_0 is rejected when it's really true
- Probability of making a type I error is denoted α and called the *level* of the test
- We choose the probability of making a type I error before conducting the test; it's often chosen to be 0.05.

Type II error

- Type II error occurs when H_0 is not rejected when it should be (when H_A is true)
- Probability of making a type II error is denoted β
- This probability is *not* determined beforehand

It isn't possible to reduce the probability of both types of error to 0 at the same time.

Example: medical treatments

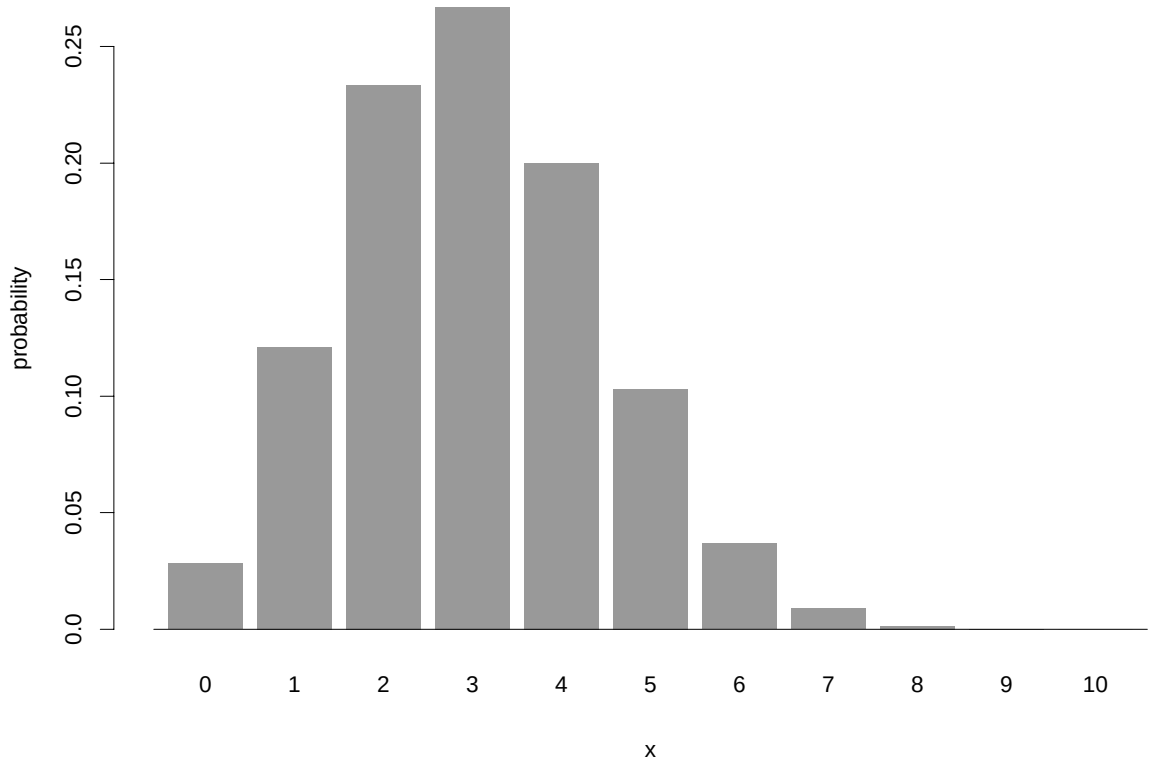
Experience has shown that 30% of people with a certain illness recover. We have developed a new medication, and we have given it to 10 randomly selected patients to test whether it increases the recovery rate.

1. What are the hypotheses that we want to test?

2. How many of them have to recover before we believe that the medicine increases the recovery rate?
   To answer this question, we want to look at the probabilities of 0, 1, 2, ... 10 patients recovering if the probability of recovery is unchanged by the new medicine.

3. If we assume that each person's probability of recovery is the same, the number of people recovering has binomial distribution with $n = 10$ and $p = 0.3$.

Binomial probabilities for our example



Larger example

- Our previous example was a very small one, but the same ideas hold for more complicated problems
- We usually have a much larger sample size, and we use the central limit theorem to approximate the sampling distributions with the normal distribution (or t distribution)

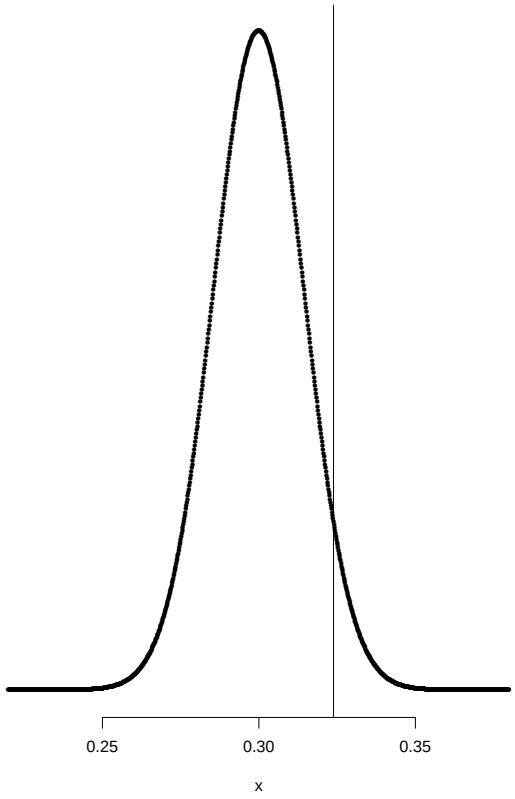


Figure 1: Approximate sampling distribution with n=1000

Establishing the rejection region

- In general, we are given α (the significance level of the test)
- We have to establish a rejection region (RR) based on α
- This means we want to find a RR such that the probability of observing an estimator in that region, given that the null hypothesis is true, is α
- We need to consider the sampling distribution of the estimator, and determine the values that fall in this range

What is a test statistic?

- Depending on H_0 the center and spread of the sampling distribution may change
- It's easier to rescale the sampling distribution to the standard normal (or t) distribution than to look at a different sampling distribution for each example
- This means that estimator ($\bar{x}$, $\hat{p}$, etc.) we observe from our data needs to be thought of in terms of how far it deviates from its expected value (under H_0)
- This re-adjusted value is called the test statistic: $\frac{\hat{\theta}-\theta_0}{SE_{\hat{\theta}}}$
- We can define the *rejection region* using standard normal (or t) distribution as well

Example: medical treatments (cont.)

Instead of randomly choosing 10 ill patients, we were able to randomly choose 1000. We observed that 350 of the 1000 recovered after being given the medicine.

Is there evidence to lead us to believe that those given the new medicine have a higher rate of recovery than 30%? Test with significance level $\alpha = 0.05$.

Example: bottling factory

A bottling factory fills thousands of 20oz bottles daily with soda, but not all the bottles are filled to the same level. A random sample of bottles was taken from the factory line, containing the following amounts of soda (in oz):

19.8 20.1 19.7 19.2 19.9 20.0 19.8 19.9 19.7

Assuming that the distribution of amounts of soda is approximately normal, is there evidence to lead us to believe that the mean level of filling is not 20 oz? Test with significance level $\alpha = 0.05$.

From the sample data, we have: $n = 9$, $\bar{x} \approx 19.79$, $s \approx 0.26$

Attained levels of significance (p-values)

- If we assume that H_0 is true, we can calculate the probability of seeing a test statistic that is this far away from what we would expect given H_0, or further
- This probabilitly is the *p-value* or *attained level of significance*
- If the p-value is less than α, this is equivalent to saying that the test statistic falls in the rejection region
- Reporting the p-value lets each reader know whether the null hypothesis would be rejected using his/her own choice of α.

Example: Comparing mean incomes

We want to compare the mean family income in two states. For state 1, we had a random sample of $n_1 = 100$ families with a sample mean of $\bar{x}_1 = 35000$. For state 2, we had a random sample of $n_2 = 144$ families with a sample mean of $\bar{x}_2 = 36000$. Past studies have shown that for both states $\sigma = 4000$.

Is the mean income in state 1 lower than that in state 2? Test at significance level $\alpha = 0.01$. Give the p-value.

Summary: Hypothesis testing methodology

1. Establish the null hypothesis.
2. Determine the alternative hypothesis $\rightarrow$ one-tailed or two-tailed test
3. Define assumptions needed for the test, make sure they are met. (Ex: paired/unpaired, equal variances, etc.)
4. Calculate the test statistic:
   $Z/T = \dfrac{\text{estimator for parameter} - \text{value for parameter given by null hypothesis}}{\text{standard error of estimator}}$
5. Based on α and H_A, determine the rejection region. This is the portion of the tail(s) (based on H_A) of the estimator's sampling distribution which has area α.
6. Determine the p-value (attained level of significance). This is the probability of seeing a result as extreme or more extreme than your test statistic under the null hypothesis, where "extreme" is defined by H_A
7. Reject H_0 if $p < \alpha$ or the test statistic falls in the rejection region

Example: Diff. in means (small samples)

The strength of concrete depends, to some extent, on the method used for drying it. Two different drying methods yielded the results below for independently tested specimens (in psi). Do the methods appear to produce concrete with a statistically significant difference in mean strengths? Use $\alpha = 0.05$. Find the attained significance level.

Method 1	$n_1 = 7$	$\bar{y}_1 = 3250$	$s_1 = 210$
Method 2	$n_2 = 10$	$\bar{y}_2 = 3240$	$s_2 = 190$

Confidence intervals vs. hypothesis tests

Imagine a two-tailed test with hypotheses $H_0 : \theta = \theta_0$ and $H_A : \theta \neq \theta_0$

- For level of significance α, reject if test statistic (could be Z or T) falls in $Z < -z_{\frac{\alpha}{2}}$ or $Z > z_{\frac{\alpha}{2}}$
- This means we fail to reject H_0 when $-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}$
- We can re-write this:
  $-z_{\frac{\alpha}{2}} < \frac{\hat{\theta}-\theta_0}{SE_{\hat{\theta}}} < z_{\frac{\alpha}{2}}$
  $-\hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}} < -\theta_0 < -\hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}}$
  $\hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}} > \theta_0 > \hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}}$
  $\hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}} < \theta_0 < \hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}}$
- As long as θ_0 lies within bounds $(\hat{\theta} - z_{\frac{\alpha}{2}} SE_{\hat{\theta}}, \hat{\theta} + z_{\frac{\alpha}{2}} SE_{\hat{\theta}})$, we fail to reject
- These are confidence interval boundaries! A confidence interval can suffice for a 2-tailed test. If interval doesn't include θ_0, then we can reject H_0

Can show similar relationship for one-sided CIs and one-tailed hypothesis tests.

Ex: Compare CIs and hypothesis tests

We have a random sample of checking account balances at each of two banks, and we're interested in testing whether the mean checking account balances are different at the two banks. Use $\alpha = 0.10$.

Bank 1 $\quad n_1 = 12 \quad \bar{x}_1 = 1000 \quad s_1 = 150$

Bank 2 $\quad n_2 = 10 \quad \bar{x}_2 = 920 \quad s_2 = 120$

More about type II error

- We've discussed the probability of type I error, α, associated with hypothesis tests, but we may also be interested in the probability of type II error, β
- This is the probability of failing to reject H_0, when H_A is really correct
- Type II error is the probability that test statistic does not fall in rejection region, given that the parameter has a value categorized by H_A
- In order to calculate this value, we need to calculate the test statistic given a specific value of the parameter under H_A
- Power of the test = 1-β

Example: Probability of type II error

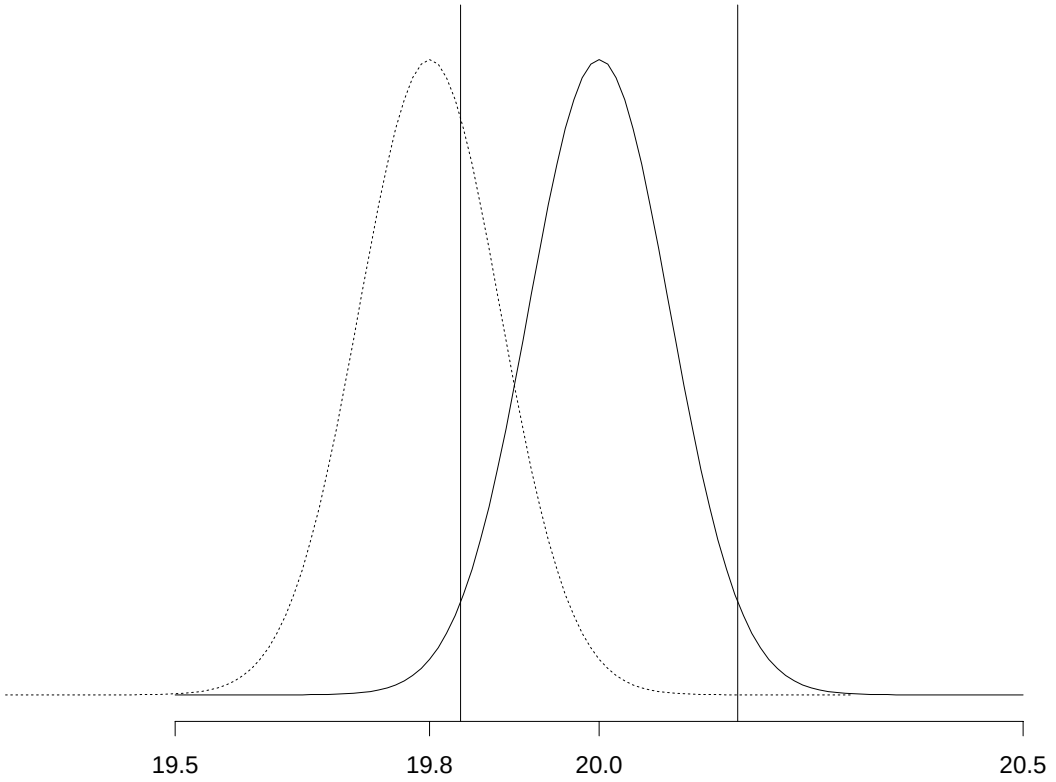
Experience has shown that 30% of people with a certain illness recover. Ten people are selected at random and given an experimental medicine; nine recover shortly thereafter.

Calculate β, the probability of type II error, with rejection region of $Y \geq 8$ and H_A: $p = 0.6$.

Example: Probability of type II error

A bottling factory fills thousands of 20oz bottles daily with soda, but not all the bottles are filled to the same level. A random sample of $n = 9$ bottles was taken from the factory line to try to determine whether the mean filling level is 20oz or not. If we test these hypotheses with $\alpha = 0.05$, find the probability of type II error if $H_A = 19.8$. Assume that the distribution of amounts of soda is approximately normal and that $\sigma = 0.25$ (known from previous experience).

Figure illustrating regions in bottling example



Example: Probability of type II error

A manufacturer claims that at least 20% of the public prefers her product. A sample of 500 people is taken to check her claim. If we test her claim with $\alpha = 0.05$, what is β if the percentage of the public prefering her product is really 22%?