6.1–6.4 Hypothesis tests

Prof. Tesler

Math 186
Winter 2019

6.1–6.2 Intro to hypothesis tests and decision rules

Hypothesis tests **are a specific way of designing experiments to quantitatively study questions like these:**

- Is a coin fair or biased? Is a die fair or biased?
- Does a gasoline additive improve mileage?
- Is a drug effective?
- Did Mendel fudge the data in his pea plant experiments?
- *Sequence alignment (BLAST):* are two DNA sequences similar by chance or is there evolutionary history to explain it?
- *DNA/RNA microarrays:*
  - Which allele of a gene present in a sample?
  - Does the expression level of a gene change in different cells?
  - Does a medication influence the expression level?

Example — Criminal trial

- In a criminal trial, the jury considers two hypotheses: innocent or guilty.
- Sometimes the evidence is clear-cut and sometimes it's ambiguous.
- **Burden of proof:** If it's ambiguous, we assume innocent. Overwhelming evidence is needed to declare guilt.
- **Mathematical language for this:**

Hypotheses

- "Null hypothesis" H_0: Innocent
- "Alternative hypothesis" H_1: Guilty

The null hypothesis, $\boldsymbol{H_0}$, is given the benefit of the doubt in ambiguous cases.

Example — Evaluating an SAT prep class

- Assume that SAT math scores are normally distributed with $\mu_0 = 500$ and $\sigma_0 = 100$.
- An SAT prep class claims it improves scores. Is it effective?
- If n people take the class, and after the class their average score is $\bar{x}$, what values of n and $\bar{x}$ would be convincing proof?

- **$\bar{x} = 502$ and $n = 10$**
  Not convincing. It's probably due to ordinary variability.
- **$\bar{x} = 502$ and $n = 1000000$**
  Convincing, although a 2 point improvement is not impressive.
- **$\bar{x} = 600$ and $n = 1$**
  Not convincing. It's just one student, who might have had a high score anyway.
- **$\bar{x} = 600$ and $n = 100$**
  Convincing.
- **$\bar{x} = 300$ and $n = 100$**
  Oops, the class made them worse!
- We need to judge these values in a quantifiable, systematic way.

Example — Evaluating an SAT prep class

Definitions

- $\mu_0 = 500$ is the average score without the class.
- μ is the theoretical average score after the class (we don't know this value however).
- $\bar{x}$ is the sample mean in our experiment (average score of our sample of students who took the class).

- If $\bar{x}$ is high, it **probably** is because the class increases scores, so the theoretical mean (μ) increased, thus increasing the sample mean ($\bar{x}$). But it's possible that the class has no effect ($\mu = \mu_0$) and we accidentally picked a sample with $\bar{x}$ unusually high.
- We assume that the scores have a normal distribution with $\sigma = \sigma_0 = 100$ with or without the class, and only consider the possibility that the class changes the mean μ.
- Later, in Chapter 7, we'll also account for changes in σ.

Hypotheses

Goal: Decide between these two hypotheses

- **"Null hypothesis":** The class has no effect.
  (Any substantial deviation of $\bar{x}$ from μ_0 is natural, due to chance.)

  H_0: $\mu = 500$ (general format: H_0: $\mu = \mu_0$)

- **"Alternative hypothesis":** The class improves the score.
  (Deviation from μ_0 is caused by the prep class.)

  H_1: $\mu > 500$ (general format: H_1: $\mu > \mu_0$)

- **Burden of proof**: Since it may be ambiguous, we assume H_0 unless there is overwhelming evidence of H_1.
- It's possible that neither hypothesis is true (for example, the distribution isn't normal; the class actually lowers the score; etc.) but the basic procedure doesn't consider that possibility.

Example — Evaluating an SAT prep class

Decision procedure (first draft)

- Pick a class of $n = 25$ people, and let $\bar{x}$ be their average score after taking the class.

  $\bar{x}$ is the *test statistic*; the decision is based on $\bar{x}$.
- If $\bar{x} \geqslant 510$, then reject H_0 (also called "reject the null hypothesis," "accept H_1," or "accept the alternative hypothesis").
- If $\bar{x} < 510$ then accept H_0 (or "insufficient evidence to reject H_0")

- The *critical region* is the values of the test statistic leading to rejecting H_0; here, it's $\bar{x} \geqslant 510$.
- The cutoff of 510 was chosen arbitrarily for this first draft. We will see its impact and how to choose a better cutoff.

Assess the error rate of this procedure

- A *Type I error* is accepting H_1 when H_0 is true.
- A *Type II error* is accepting H_0 when H_1 is true.
- First, we will focus on controlling the *Type I error rate, α*:

$$\alpha = P(\text{accept } H_1 | H_0 \text{ true}) = P(\overline{X} \geqslant 510 \mid \mu = 500)$$

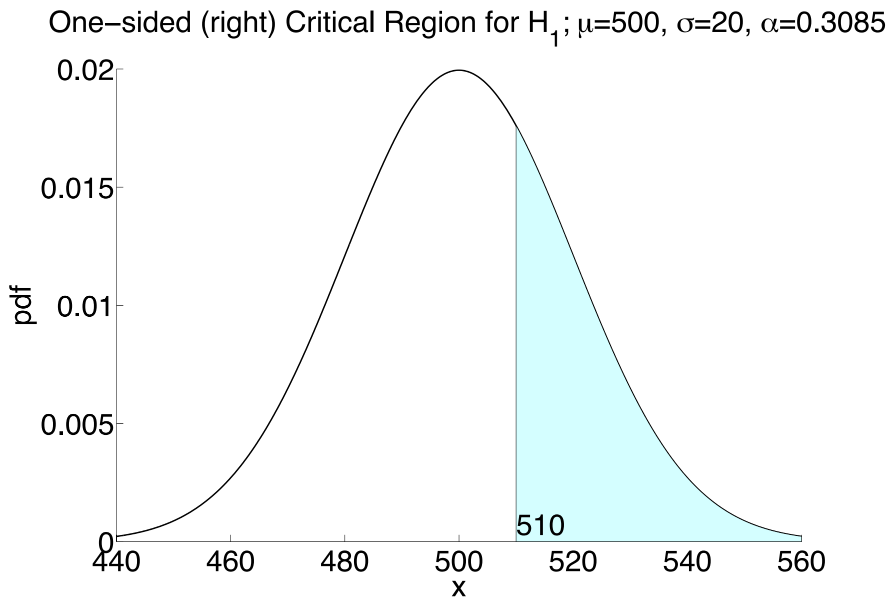
(Later, we will see how to control the Type II error rate.)

- Convert $\bar{x}$ to z-score $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \dfrac{\bar{x} - 500}{100 / \sqrt{25}}$:

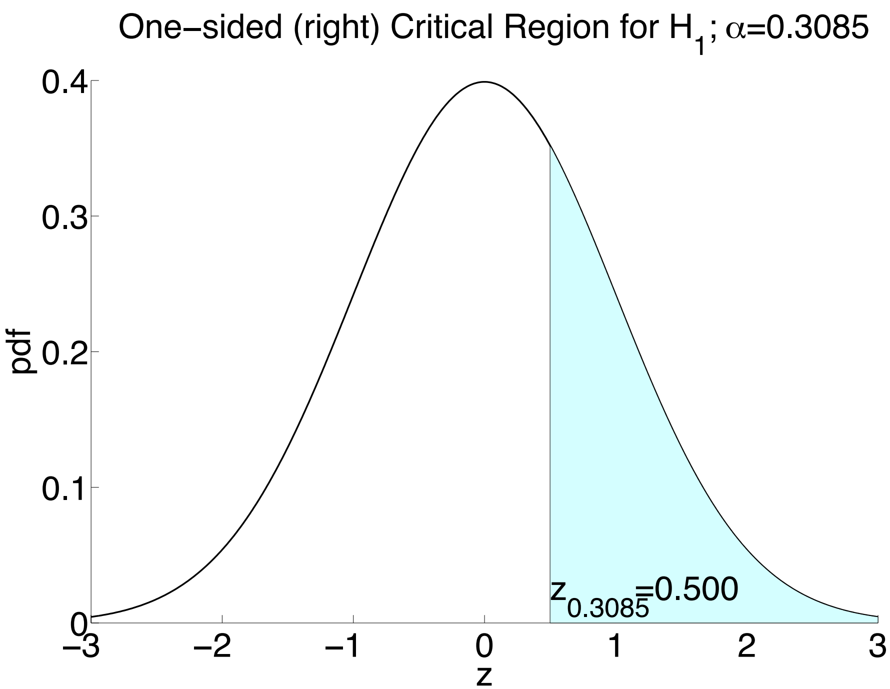
$$\begin{aligned} \alpha &= P\left(\frac{\overline{X} - 500}{100/\sqrt{25}} \geqslant \frac{510 - 500}{100/\sqrt{25}} \right) \\ &= P(Z \geqslant .5) \\ &= 1 - \Phi(.5) = 1 - .6915 = .3085 \end{aligned}$$

Critical region

Critical region in terms of $\overline{X}$



Critical region in terms of Z



- In each graph, the shaded area is $.3085 = 30.85\%$.
- When H_0 ($\mu = 500$) is true, about 30.85% of 25 person samples will have an average score $\geqslant 510$, and thus will be misclassified by this procedure.
- This test has an $\alpha = .3085$ *significance level*, which is very large.

How to choose the cutoff in the decision procedure

- Choose the *significance level*, α, first. Typically, $\alpha = 0.05$ or 0.01. Then compute the cutoff $\bar{x}$ that achieves that significance level, so that if H_0 is true, then at most a fraction α of cases will be misclassified as H_1 (a *Type I error*).
- We'll still use $n = 25$ people, but we want to find the cutoff for a significance level $\alpha = .05$.
- Solve $\Phi(z_{.05}) = .95$: $\Phi(1.64) = .95$ so $z_{.05} = 1.64$.
(For two-sided 95% confidence intervals, we used $z_{.025} = 1.96$.)
- Find the value $\bar{x}^*$ with z-score 1.64.
It's called the *critical value*, and we reject H_0 when $\bar{x} \geqslant \bar{x}^*$.

$$\frac{\bar{x}^* - 500}{100/\sqrt{25}} = 1.64$$

so

$$\bar{x}^* = 500 + 1.64 \cdot (100/\sqrt{25}) = 532.8$$

SAT prep class — Decision procedure (second draft)

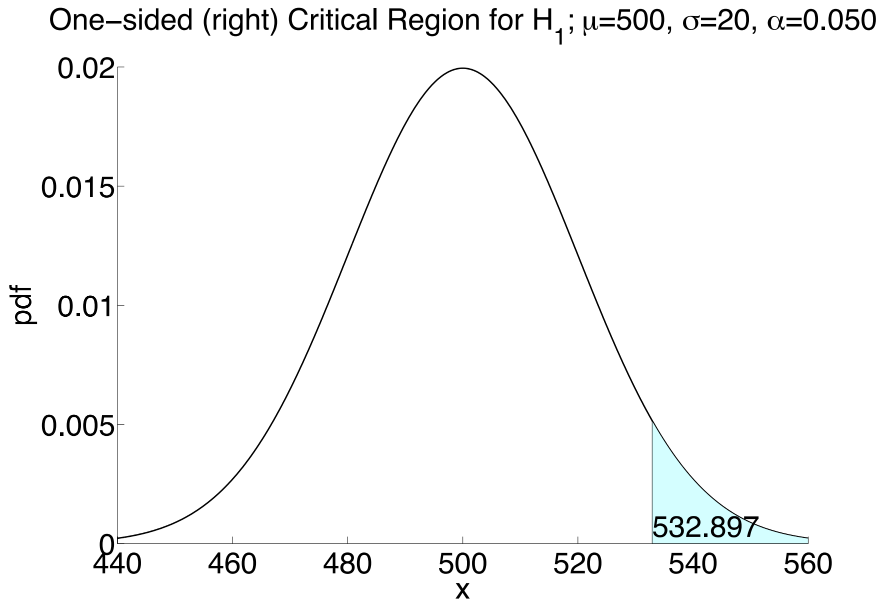
Decision procedure for 5% significance level

- Pick a class of $n = 25$ people, and let $\bar{x}$ be their average score after taking the class.
- If $\bar{x} \geqslant 532.8$ then reject H_0.
- If $\bar{x} < 532.8$ then accept H_0.

- The values of $\bar{x}$ for which we reject H_0 form the *one-sided critical region*: $[532.8, \infty)$.
- The values of $\bar{x}$ for which we accept H_0 form the *one-sided acceptance region* for μ under H_0: $(-\infty, 532.8)$.

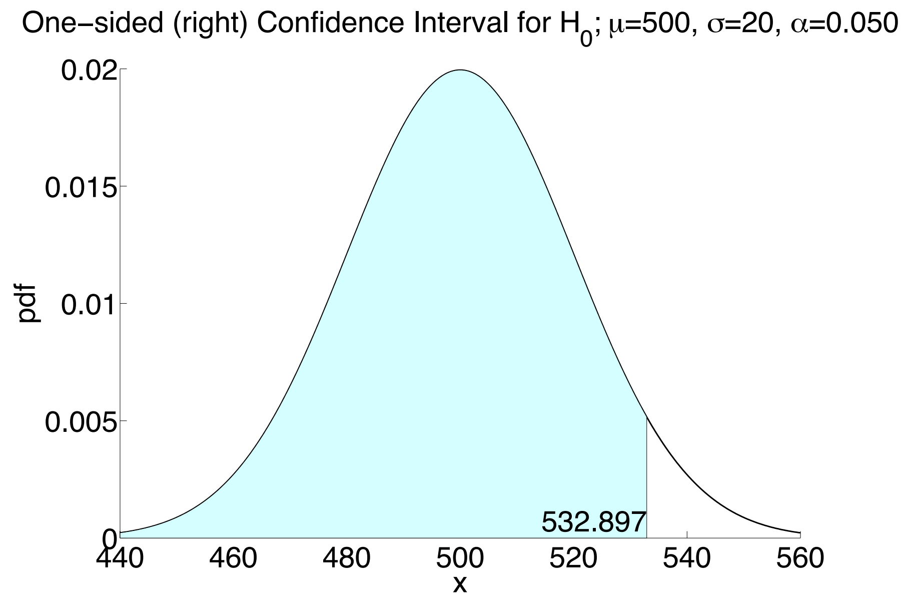
Reject H_0 if $\bar{x}$ in one-sided critical region $[532.8, \infty)$.

Area $= \alpha = .05$



Accept H_0 if $\bar{x}$ in one-sided 95% acceptance region for H_0 $(-\infty, 532.8)$.

Area $= 1 - \alpha = .95$



Type II error rate

- We designed the experiment to achieve a Type I error rate 5%.
- What is the Type II error rate (β)? For example, what fraction of the time will this procedure fail to recognize that μ rose to 530 (since that's just below 532.8)? Compute

$$\begin{aligned} \beta &= P(\text{Accept } H_0 \,|\, H_1 \text{ is true, with } \mu = 530) \\ &= P(\overline{X} < 532.8 \,|\, \mu = 530) \end{aligned}$$

- When $\mu = 530$, the z-score is *not* $\frac{\bar{x}-500}{100/\sqrt{25}}$; it's $z' = \frac{\bar{x}-530}{100/\sqrt{25}}$. So

$$\begin{aligned} \beta &= P(\overline{X} < 532.8 \,|\, \mu = 530) \\ &= P\left(\frac{\overline{X}-530}{100/\sqrt{25}} < \frac{532.8-530}{100/\sqrt{25}}\right) = P(Z' < .14) = .5557 \end{aligned}$$

- β is more complicated to define than α, because β depends on the value of the unknown parameter ($\mu = 530$ in this case), whereas for α the parameter value ($\mu = 500$) is specified in H_0.

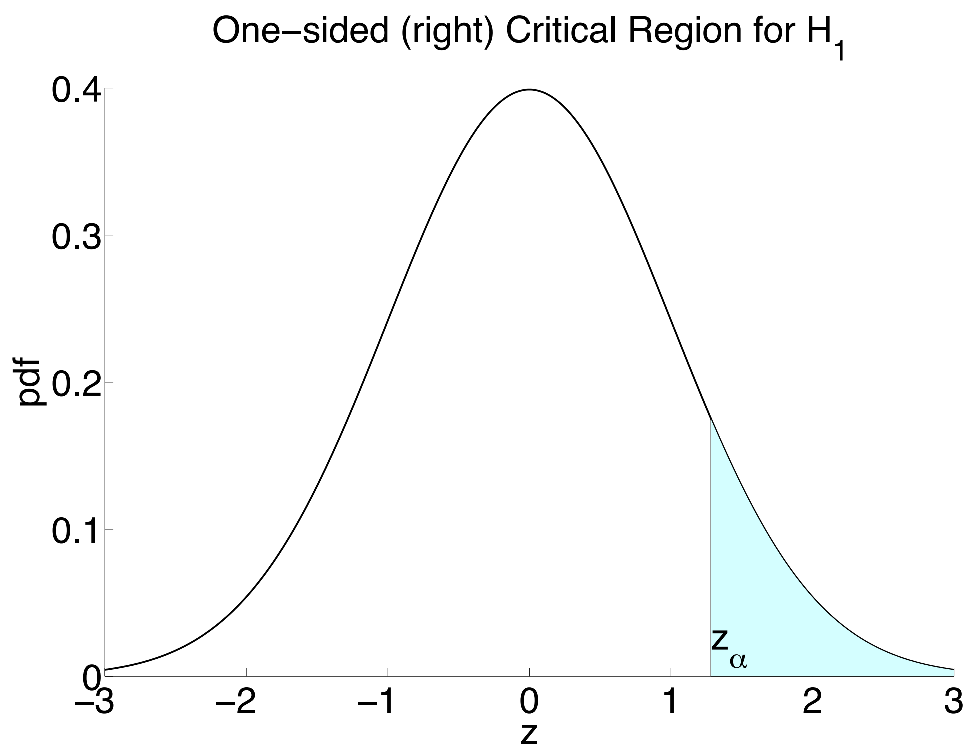
Variation (a): One-sided to the right (what we did)

Hypotheses: H_0: $\mu = 500$ vs. H_1: $\mu > 500$.

Decision: Reject H_0 if $z \geqslant z_\alpha$.
Equivalently, reject H_0 if $\bar{x} \geqslant 500 + z_\alpha \frac{\sigma}{\sqrt{n}}$.

Decision for $\alpha = 0.05$, $\sigma = 100$, $n = 25$:
Reject H_0 if $z \geqslant 1.64$.
Equivalently, reject H_0 if $\bar{x} \geqslant 500 + 1.64(\frac{100}{\sqrt{25}}) = 532.8$.

Critical region:
Gives an area α on the right.



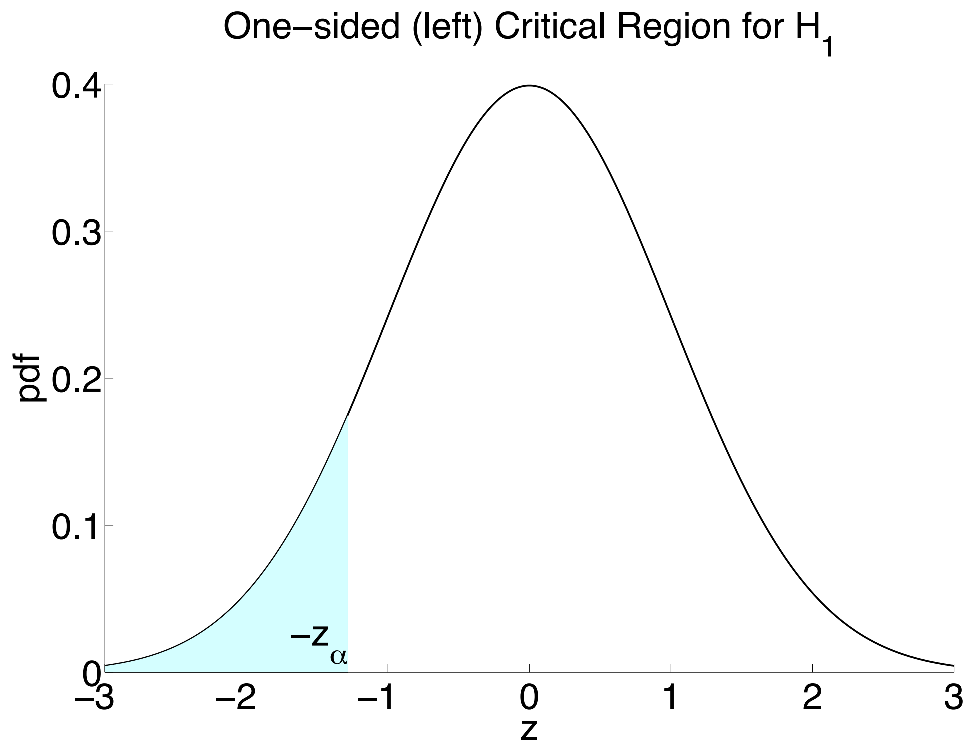
Variation (b): One-sided to the left

Hypotheses: H_0: $\mu = 500$ vs. H_1: $\mu < 500$.

Decision: Reject H_0 if $z < -z_\alpha$.
Equivalently, reject H_0 if $\bar{x} \leqslant 500 - z_\alpha \frac{\sigma}{\sqrt{n}}$.

Decision for $\alpha = 0.05$, $\sigma = 100$, $n = 25$:
Reject H_0 if $z \leqslant -1.64$.
Equivalently, reject H_0 if $\bar{x} \leqslant 500 - 1.64(\frac{100}{\sqrt{25}}) = 467.2$.

Critical region:
Gives an area α on the left.



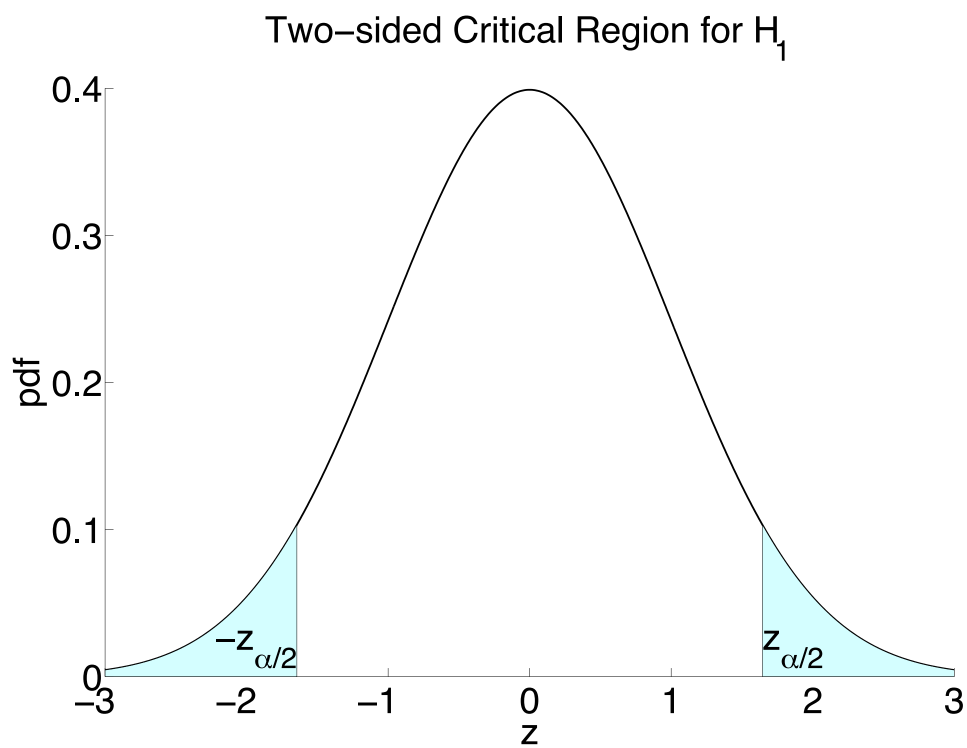
Variation (c): Two-sided

Hypotheses: H_0: $\mu = 500$ vs. H_1: $\mu \neq 500$.

Decision: Reject H_0 if $|z| \geqslant z_{\alpha/2}$.
Equivalently, reject H_0 unless $\bar{x}$ is between $500 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Decision for $\alpha = 0.05$, $\sigma = 100$, $n = 25$:
Reject H_0 if $|z| \geqslant 1.96$. Equivalently,
reject H_0 unless $\bar{x}$ is between $500 \pm 1.96\frac{100}{\sqrt{25}} = (460.8, 539.2)$

Critical region:
Gives an area α split up as $\alpha/2$ on each side.



Variations — Summary

(a) For H_0: $\mu = 500$ vs. H_1: $\mu > 500$,
the critical region is an area $\alpha = 5\%$ at the right.

(b) For H_0: $\mu = 500$ vs. H_1: $\mu < 500$,
the critical region is an area $\alpha = 5\%$ at the left.

(c) For H_0: $\mu = 500$ vs. H_1: $\mu \neq 500$,
the critical region is split into area $\alpha/2 = 2.5\%$ at the right and $\alpha/2 = 2.5\%$ at the left.

- "500" and "5%" can be replaced by other constant values.
- Important values of z_α (look up others in the table in the book):

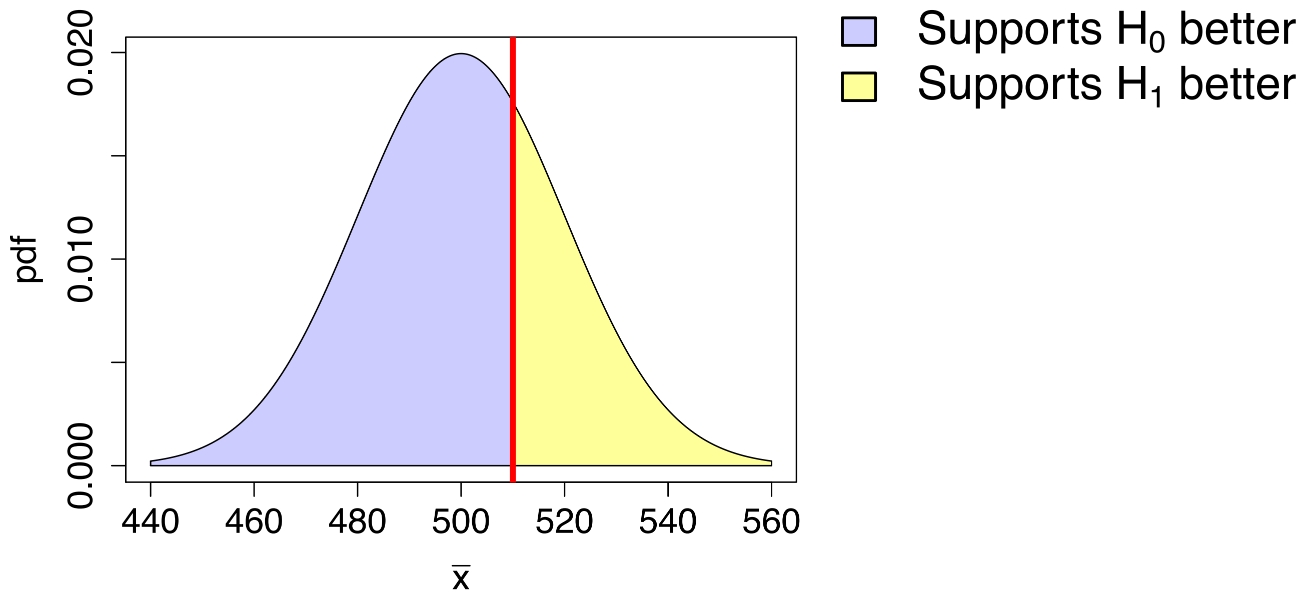
	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
One-sided	$z_{.01} \approx 2.33$	$z_{.05} \approx 1.64$	$z_{.10} \approx 1.28$
Two-sided	$z_{.005} \approx 2.58$	$z_{.025} \approx 1.96$	$z_{.05} \approx 1.64$

P-values

- Another way to do hypothesis tests. Makes the same conclusions.
- A Type I error is accepting H_1 when H_0 is really true.
- This happens because we got an unusually bad sample, where the test statistic accidentally falls in the critical region.
- Given a sample with a particular test statistic, its *P-value* is the probability to draw another sample with an even worse test statistic (meaning more supportive than the current sample of making the incorrect decision "Accept H_1" / "Reject H_0").

P-values

Consider H_0: $\mu = 500$ vs. H_1: $\mu > 500$ with $\sigma = 100$ and $n = 25$



- Suppose our sample has $\bar{x} = 510$.
- Samples supporting H_1 / opposing H_0 as much or more than this one are those with $\bar{x} \geqslant 510$.
- We showed $\bar{x} \geqslant 510$ for $\approx 30.85\%$ of all samples when H_0 is true:

$$P(\overline{X} \geqslant 510 | H_0) = P\left(\frac{\overline{X} - 500}{100/\sqrt{25}} \geqslant \frac{510-500}{100/\sqrt{25}}\right)$$
$$= P(Z \geqslant .5) = 1 - \Phi(.5) = 1 - .6915 = .3085$$

- The *P-value* of $\bar{x} = 510$ is $P = .3085 = 30.85\%$.

P-values

Consider H_0: $\mu = 500$ vs. H_1: $\mu > 500$ with $\sigma = 100$ and $n = 25$

- This means the probability under H_0 of seeing a value "at least as extreme" as $\bar{x} = 510$ is 30.85%.
- For other decision procedures, the definition of *"at least this extreme"* (more supportive of H_1, less supportive of H_0) depends on the hypotheses.
- The z-score of $\bar{x} = 510$ under H_0 is $z = \frac{510-500}{100/\sqrt{25}} = \frac{10}{20} = .5$.
  H_1 says what it means to be at least that extreme:

(a) H_0: $\mu = 500$ vs. H_1: $\mu > 500$.

$$P = P(\overline{X} \geqslant 510) = P(Z \geqslant .5) = 1 - \Phi(.5) = 1 - .6915 = .3085$$

(b) H_0: $\mu = 500$ vs. H_1: $\mu < 500$.

$$P = P(\overline{X} \leqslant 510) = P(Z \leqslant .5) = \Phi(.5) = .6915$$

(c) H_0: $\mu = 500$ vs. H_1: $\mu \neq 500$.

$$P = P(\overline{X} \geqslant 510) + P(\overline{X} \leqslant 490)$$

$$= P(|Z| \geqslant .5) = P(Z \geqslant .5) + P(Z \leqslant -.5) = .3085 + .3085 = .6170$$

P-values for $\bar{x} = 510$ ($z = .5$) for different H_1's

(a) H_0: $\mu = 500$
H_1: $\mu > 500$

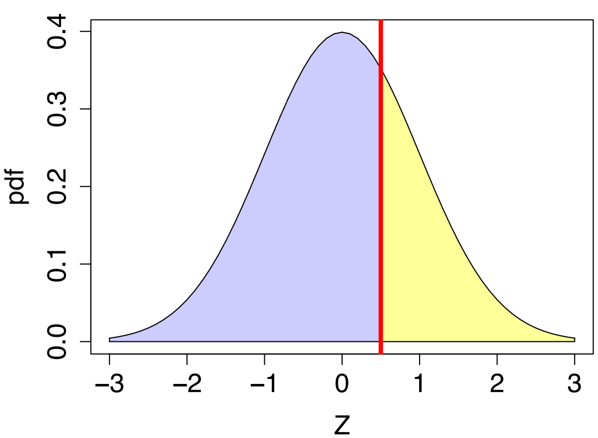
$$\begin{aligned} P &= P(Z \geqslant .5) \\ &= 1 - \Phi(.5) \\ &= 1 - .6915 \\ &= .3085 \end{aligned}$$

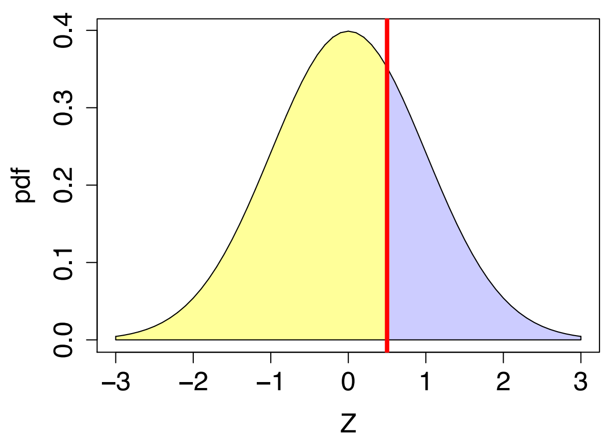
(b) H_0: $\mu = 500$
H_1: $\mu < 500$

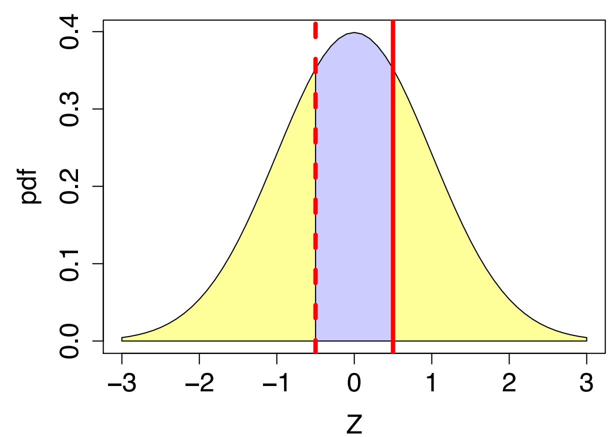
$$\begin{aligned} P &= P(Z \leqslant .5) \\ &= \Phi(.5) \\ &= .6915 \end{aligned}$$

(c) H_0: $\mu = 500$
H_1: $\mu \neq 500$

$$\begin{aligned} P &= P(|Z| \geqslant .5) \\ &= 2P(Z \geqslant .5) \\ &= 2(.3085) \\ &= .6170 \end{aligned}$$







- Supports H_0 better
- Supports H_1 better
- Observed z=0.50

P-values

- In terms of P-values, the decision procedure is

  "Reject H_0 if $P \leqslant \alpha$."

- **Interpretation:** Suppose $P \leqslant \alpha$. If H_0 holds, events at least this extreme are rare, occurring $\leqslant (100\alpha)$% of the time. But if H_1 holds, there's a much higher probability of test statistics in this range. Since we observed this event, H_1 is more plausible.
  (a) $P{=}0.3085$. When H_0 holds, about 30.85% of samples have $\overline{X} \geqslant 510$.
  (b) $P{=}0.6915$. When H_0 holds, about 69.15% of samples have $\overline{X} \leqslant 510$.
  (c) $P{=}0.6170$. When H_0 holds, about 61.70% of samples have either $\overline{X} \geqslant 510$ or $\overline{X} \leqslant 490$.

- At the $\alpha = .05$ significance level, we accept H_0 in all three cases since $P > .05$. Events this "extreme" are very common under H_0, so this does not provide convincing evidence against H_0.

P-values for $\bar{x} = 536$

- Suppose $n = 25$ and $\bar{x} = 536$.
- Then $z = \frac{536-500}{100/\sqrt{25}} = \frac{36}{20} = 1.8$

(a) H_0: $\mu = 500$ vs. H_1: $\mu > 500$

- The P-value is $P = P(Z \geqslant 1.8) = 1 - \Phi(1.8) = 1 - .9641 = .0359$.
- If H_0 is true, only 3.59% of the time would we get a score this extreme or worse.
- At $\alpha = .05$, we reject H_0, since $P \leqslant \alpha$: $.0359 \leqslant .05$.
- At $\alpha = .01$, we accept H_0 since $P > \alpha$: $.0359 > .01$.
  Another interpretation is we do not have sufficient evidence to reject H_0 at significance level $\alpha = .01$.

P-values for $\overline{X} = 536$

- Suppose $n = 25$ and $\overline{X} = 536$.
- Then $z = \frac{536-500}{100/\sqrt{25}} = \frac{36}{20} = 1.8$

(c) H_0: $\mu = 500$ vs. H_1: $\mu \neq 500$

- The P-value is $P = P(|Z| \geqslant 1.8) = 2(.0359) = .0718$
- Accept H_0 at both $.01$ and $.05$ significance levels since $.0718 > .01$ and $.0718 > .05$.

Advantages of P-values over critical values for hypothesis tests

- P-values give a continuous scale, so if you're near the arbitrary cutoff, you know it.
- P-values allow you to test against cutoffs for several α's simultaneously. We could compute the critical values of $\bar{x}$ for $\alpha = 0.01, 0.05$, etc., but this saves some steps.
- P-values can be defined for any statistical distribution, not just the normal distribution, so hypothesis tests for any distribution can be formulated as "Reject H_0 if $P \leqslant \alpha$."
- You can pick up a scientific paper that uses any statistical distribution, even a distribution you don't yet know, and still understand the results if they are expressed using P-values. Otherwise, for each new test statistic, you have to learn the details of the test and how to interpret the test statistic.

Consider a coin with probability p of heads, $1 - p$ of tails.
Warning: do not confuse this with the P from P-values.

Two-sided hypothesis test: Is the coin fair?

Null hypothesis: H_0: $p = .5$ ("coin is fair")
Alternative hypothesis: H_1: $p \neq .5$ ("coin is not fair")

Draft of decision procedure

- Flip a coin 100 times.
- Let X be the number of heads.
- If X is "close" to 50 then it's fair, and otherwise it's not fair.

How do we quantify "close"?

Decision procedure — confidence interval

How do we quantify "close"?

Form a 95% confidence interval for the expected # of heads:

$n = 100$, $p = 0.5$

$$\mu = np = 100(.5) = 50$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{100(.5)(1-.5)} = \sqrt{25} = 5$$

Using the normal approximation, the 95% confidence interval is

$$\begin{aligned}(\mu - 1.96\sigma, \mu + 1.96\sigma) &= (50 - 1.96 \cdot 5\ ,\ 50 + 1.96 \cdot 5) \\ &= (40.2\ ,\ 59.8)\end{aligned}$$

Check that it's OK to use the normal approximation

$$\mu - 3\sigma = 50 - 15 = 35 > 0$$
$$\mu + 3\sigma = 50 + 15 = 65 < 100$$ so it is OK.

Decision procedure

Hypotheses

Null hypothesis: H_0: $p = .5$ ("coin is fair")
Alternative hypothesis: H_1: $p \neq .5$ ("coin is not fair")

Decision procedure

- Flip a coin 100 times.
- Let X be the number of heads.
- If $40.2 < X < 59.8$ then accept H_0; otherwise accept H_1.

Significance level: $\approx 5\%$

If H_0 is true (coin is fair), this procedure will give the wrong answer (H_1) about 5% of the time.

Measuring Type I error (a.k.a. *Significance Level*)

H_0 is the true state of nature, but we mistakenly reject H_0 / accept H_1

- If this were truly the normal distribution, the Type I error would be $\alpha = .05 = 5\%$ because we made a 95% confidence interval.
- However, the normal distribution is just an approximation; it's really the binomial distribution. So:

$$\begin{aligned}
\alpha &= P(\text{accept } H_1 | H_0 \text{ true}) \\
&= 1 - P(\text{accept } H_0 | H_0 \text{ true}) \\
&= 1 - P(40.2 < X < 59.8 \,|\, \text{binomial with } p = .5) \\
&= 1 - .9431120664 = 0.0568879336 \approx 5.7\%
\end{aligned}$$

$$\begin{aligned}
P(40.2 < X < 59.8 \,|\, p = .5) &= \sum_{k=41}^{59} \binom{100}{k} (.5)^k (1 - .5)^{100-k} \\
&= .9431120664
\end{aligned}$$

- So it's a 94.3% confidence interval and the Type I error rate is $\alpha = 5.7\%$.

Measuring Type II error

H_1 is the true state of nature but we mistakenly accept H_0 / reject H_1

- If $p = .7$, the test will probably detect it.
- If $p = .51$, the test will frequently conclude H_0 is true when it shouldn't, giving a high Type II error rate.
- If this were a game in which you won \$1 for each heads and lost \$1 for tails, there would be an incentive to make a biased coin with p just above $.5$ (such as $p = .51$) so it would be hard to detect.

Measuring Type II error

Exact Type II error for $p = .7$ using binomial distribution

- $$\begin{aligned}\beta &= P(\text{Type II error with } p = .7)\\ &= P(\text{Accept } H_0 \,|\, X \text{ is binomial, } p = .7)\\ &= P(40.2 < X < 59.8 \,|\, X \text{ is binomial, } p = .7)\\ &= \sum_{k=41}^{59} \binom{100}{k} (.7)^k (.3)^{100-k} = .0124984 \approx 1.25\%.\end{aligned}$$

- **When $p = 0.7$, the Type II error rate, β, is $\approx 1.25\%$:** $\approx 1.25\%$ of decisions made with a biased coin (specifically biased at $p = 0.7$) would incorrectly conclude H_0 (the coin is fair, $p = 0.5$).

- Since H_1: $p \neq .5$ includes many different values of p, the Type II error rate depends on the specific value of p.

Measuring Type II error

Approximate Type II error using normal distribution

- $\mu = np = 100(.7) = 70$
- $\sigma = \sqrt{np(1-p)} = \sqrt{100(.7)(.3)} = \sqrt{21}$
- $$\begin{aligned} \beta &= P(\text{Accept } H_0 \,|\, H_1 \text{ true: } X \text{ binomial with } n = 100,\ p = .7) \\ &\approx P(40.2 < X < 59.8 \,|\, X \text{ is normal with } \mu = 70,\ \sigma = \sqrt{21}) \\ &= P\left(\frac{40.2-70}{\sqrt{21}} < \frac{X-70}{\sqrt{21}} < \frac{59.8-70}{\sqrt{21}}\right) \\ &= P(-6.50 < Z < -2.23) \\ &= \Phi(-2.23) - \Phi(-6.50) \\ &= .0129 - .0000 = .0129 = 1.29\% \end{aligned}$$

  which is close to the correct value $\approx 1.25\%$ that we found by summing the binomial distribution.
- There are also rounding errors from using the table in the book instead of a calculator that computes $\Phi(z)$ more precisely.

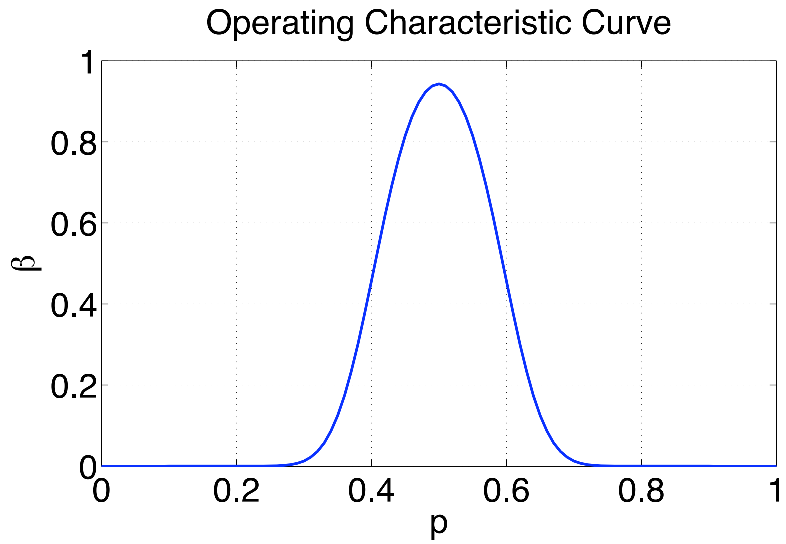
Power curve

- The decision procedure is "Flip a coin 100 times, let X be the number of heads, and accept H_0 if $40.2 < X < 59.8$".
- Plot the Type II error rate as a function of p:

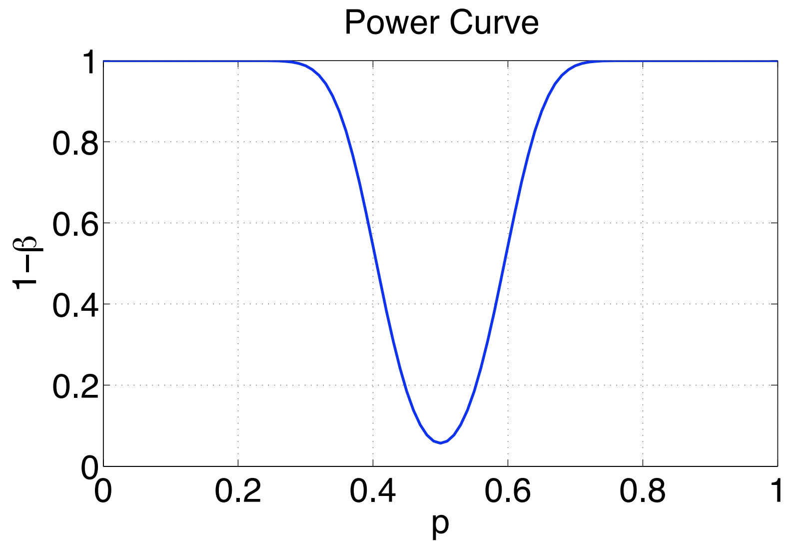
$$\beta = \beta(p) = \sum_{k=41}^{59} \binom{100}{k} p^k (1-p)^{100-k}$$

Type II Error:

$\beta = P($**Accept** $H_0 \mid H_1$ **true**$)$



Correct detection of H_1:
Power = Sensitivity =
$1 - \beta = P($**Accept** $H_1 \mid H_1$ **true**$)$



Choosing n to control Type I and II errors together

- Suppose we increase α from 0.05 to 0.10.
  - All samples with P-values between 0.05 and 0.10 are reclassified from **Accept H_0** into **Reject H_0.**
  - Samples with any other P-values are classified the same as before.
  - Thus, increasing α increases the Type I error rate and decreases the Type II error rate. Decreasing α does the reverse.
- To keep both Type I & Type II errors down, we need to increase n.
- For a null hypothesis H_0: $p = 0.50$, we want a test that is able to detect $p = 0.51$ at the $\alpha = 0.05$ significance level.

Choosing n to control Type I and II errors together

Goal: Detect $p = 0.51$ when $p = 0.50$ is supposed to hold

- For $n = 100$, it's hard to distinguish $p = 0.50$ from 0.51, since the intervals supporting those are nearly the same, while for $n = 1$ million, there's no overlap (all for $\alpha = 0.05$):

$\boldsymbol{p}$	**2-sided acceptance interval for** $\boldsymbol{n = 100}$	$\boldsymbol{n = 1}$ **million**
$p = 0.50$	$k = 41, \cdots, 59$	$k = 499020, \cdots, 500980$
$p = 0.51$	$k = 42, \cdots, 60$	$k = 509021, \cdots, 510979$

- We'll see how to compute what n to use instead of just guessing a big number.

- Also, our goal is to detect an increase in p, so it's better to use a 1-sided test instead of a 2-sided test.

Choosing n to control Type I and II errors together

Goal: Detect $p = 0.51$ when $p = 0.50$ is supposed to hold

General format of hypotheses for p in a binomial distribution

$$H_0\colon p = p_0$$

vs. one of these for H_1:

$$H_1\colon p > p_0$$
$$H_1\colon p < p_0$$
$$H_1\colon p \neq p_0$$

where p_0 is a specific value.

Our hypotheses

$$H_0\colon p = .5 \quad \text{vs.} \quad H_1\colon p > .5$$

Choosing n to control Type I and II errors together

Hypotheses

$$H_0\colon p = .5 \qquad \text{vs.} \qquad H_1\colon p > .5$$

Analysis of decision procedure

- Flip the coin n times, and let x be the number of heads.
- Under the null hypothesis, $p_0 = .5$ so
$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{x - .5n}{\sqrt{n(.5)(.5)}} = \frac{x - .5n}{\sqrt{n}/2}$$
- The z-score of $x = .51n$ is $z = \dfrac{.51n - .5n}{\sqrt{n}/2} = .02\sqrt{n}$
- We reject H_0 when $z \geqslant z_\alpha = z_{0.05} = 1.64$, so

$$.02\sqrt{n} \geqslant 1.64 \qquad \sqrt{n} \geqslant \frac{1.64}{.02} = 82 \qquad n \geqslant 82^2 = 6724$$

Choosing n to control Type I and II errors together

- Thus, if the test consists of $n = 6724$ flips, only $\approx 5\%$ of such tests on a fair coin would give $\geqslant 51\%$ heads.

- Increasing n further reduces the fraction α of tests giving $\geqslant 51\%$ heads with a fair coin.

- Instead of using the number of heads x, we could have used the proportion of heads $\hat{p} = \bar{x} = x/n$, which gives z-score

$$z = \frac{(x/n) - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{(x/n) - .5}{1/(2\sqrt{n})} = \frac{x - .5n}{\sqrt{n}/2}$$

which is the same as before, so the rest works out the same.

Sec. 6.4. Errors in hypothesis testing

Terminology: Type I or II error

Decision	True state of nature: H_0 true	True state of nature: H_1 true
Accept H_0 / Reject H_1	Correct decision	Type II error
Reject H_0 / Accept H_1	Type I error	Correct decision

Alternate terminology:

Null hypothesis H_0=“negative”
Alternative hypothesis H_1=“positive”

Decision	True state of nature: H_0 true	True state of nature: H_1 true
Acc. H_0 / Rej. H_1 / “negative”	True Negative (TN)	False Negative (FN)
Rej. H_0 / Acc. H_1 / “positive”	False Positive (FP)	True Positive (TP)

Measuring α and β from empirical data

Suppose you know the # times the tests fall in each category

Decision	True state of nature: H_0 true	H_1 true	Total
Accept H_0 / Reject H_1	1	2	3
Reject H_0 / Accept H_1	4	10	14
Total	5	12	17

Error rates

Type I error rate: $\alpha = P(\text{reject } H_0 | H_0 \text{ true}) = 4/5 = .8$

Type II error rate: $\beta = P(\text{accept } H_0 | H_0 \text{ false}) = 2/12 = 1/6$

Correct decision rates

Specificity: $1 - \alpha = P(\text{accept } H_0 | H_0 \text{ true}) = 1/5 = .2$

Sensitivity: $1 - \beta = P(\text{reject } H_0 | H_0 \text{ false}) = 10/12 = 5/6$

Power = sensitivity = $5/6$

Errors in hypothesis testing

- Type I and II errors assume that one of them is right and analyze the probabilities of choosing the wrong one.
- The theoretical analysis assumes we know the correct probability distribution. It's best to check this, e.g., by making a histogram of tons of data.
- For coin flips, the binomial distribution is the right model.
- SATs and other exam scores are often assumed to follow a normal distribution, but it may not be true.

Mendel's Pea Plant Experiments

Mendel observed 7 traits in his pea plant experiments. He determined the genotype for tall/short as follows (and the other traits were done in an analagous way):

Mendel's Decision Procedure

- If a plant is short, its genotype is tt.
- If a plant is tall, do an experiment to determine if the genotype is Tt or TT: self-fertilize the plant, get 10 seeds, and plant them.
  - If any of the offspring are short, the original plant is declared to have genotype Tt (heterozygous).
  - If all offspring are tall, the original plant is declared to have genotype TT (homozygous).

Mendel's Pea Plant Experiments

- If this procedure gives tt or Tt, it's correct.
- However, classifications as TT might be erroneous!
- Assuming the genotypes of separate offspring are independent, if the original plant is heterozygous (Tt), the probability of it producing 10 tall offspring is

$$(.75)^{10} = .05631351$$

- Thus, about 5.6% of Tt plants will be incorrectly classified as TT.
- When tall plants are tested relative to the hypotheses

  H_0: *genotype is* Tt *vs.* H_1: *genotype is TT*

  the Type I error rate is $\alpha \approx .056$ and the Type II error rate is $\beta = 0$.