Ch 10 Hypothesis Testing

Motivational Example: I claim I am a 75% free throw shooter.

To test, you have me shoot 20 f.t.

I make 8.

You claim this proves I am not a 75% f.t. shooter.

Are you right?

We can think of this as a binomial experiment using $p = 0.75$.

$$P(T=8) = \binom{20}{8}(3/4)^8(1/4)^{12}$$
$$= 0.00075.$$

Well, all that means is we observed a rare event if p is true.

But another way to think about it is, if $p = 0.75$ what is the likelihood that I would ever shoot 8 or worse.

$$P(T \le 8) = \sum_0^8 \binom{20}{n}(0.75)^n(0.25)^{20-n}$$

< 0.001 !

The entire area is less than $\frac{1}{1000}$!

We a probably safe to concluded that my true percentage $p < 0.75$.

This is what hypothesis testing is all about.

- assume something is true.
- observe and compute the prob. of seeing this outcome or more extreme.
- decide if there is evidence against your assumption.

Flip-side.

In Ch8 we would have constructed a confidence interval.

$n = 20$ is small ... need $t(19)$ dist'n.

A 99% confidence interval for p uses $t_{0.005}(19) = 2.861$

$$\hat{p} \pm t_{0.005}\sqrt{\frac{\hat{p}\hat{q}}{20}}, \quad 0.4 \pm 0.256$$

or $(0.144, 0.656)$

$p = 0.75$ way outside of this.

Again, very unlikely $p = 0.75$.

* Hypo. Testing & C.I. are 2 sides of the same coin.

§ 10.2. 10.3

disc: Hypo. Testing Vocab.

① the hypotheses.

- the null hypothesis H_0 - status quo.

  * assumed to be true unless "proven" otherwise

  eg. $H_0: p = 0.75$

- the alternative hypothesis H_a

  this is the claim we aim to test

  eg. $H_a: p < 0.75$.

② the decision rule

this is the cutoff α we use to accept or reject H_a.

usually stated beforehand

③ the decision

after a probability calculation. either

(i) reject the null hypo. H_0

(ii) not enough evidence to reject H_0.

④ p-value.

The probability of seeing as much or more evidence for H_a than we saw in the data.

eg. p-value < 0.001.

Smaller p-value, more evidence for H_a.

(*) test results are called <u>statistically significant</u> if H_0 is rejected.

topic: Error Types.

2 possible decisions $\Rightarrow$ 2 possible mistakes:

	The "truth"	H_0	H_a
Our test supports	H_0	✓	Type II
	H_a	Type I	✓ good

Type I: we reject H_0 when it is true.
this error comes from our decision rule.
e.g. I actually am 75% FT shooter, but had a bad day.

Type II: we don't reject H_0 when H_a is true.

We will see that we can measure these error is some sense.

$$P(\text{Type I error}) = \alpha$$
$$P(\text{Type II error}) = \beta$$

ex: back to the free throws.

Before hand, you decide that if I make 10 or less, you will reject $H_0: p = 75\%$

this is the decision rule.

The text calls this set of outcomes the rejection region (RR).

If $T \in \{0, 1, 2, \ldots, 10\}$, this supports H_a.

We can calculate the probability of this

$$
\begin{aligned}
\alpha &= P(\text{Type I error}) \\
&= P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \\
&= P(0 \le T \le 10 \text{ when } p = 0.75) \\
&= \text{pbinom}(10, 20, 0.75) \\
&= 0.013.
\end{aligned}
$$

Computing Type II error requires a guess for H_a.
In our example, need to use a $p < 3/4$.

$$
\begin{aligned}
\beta &= P(\text{Type II error}) \\
&= P(\text{accept } H_0 \text{ when } H_a \text{ is true}) \\
&= P(11 \le T \le 20 \text{ when } p = \theta < 3/4) \\
&= 1 - \text{pbinom}(10, 20, \theta)
\end{aligned}
$$

$\beta(\theta)$, a fcn of θ.

Here $\theta = 0.6$, $\beta = 0.755$
(that is a lot)

If $\theta = 0.5$, $\beta = 0.4119$.

Note the larger the true difference btwn true $p = \theta$ and the null hypothesis $p_0 = 0.75$, the smaller the Type II error.

Of course increasing the rejection region will increase α but decrease β.

def. the power fcn of the test is defined
$$\text{Power}(\theta) = 1 - \beta(\theta).$$

* this measures Type II error.

ex: new ultrasound machine.
Claims to detect tumors better than older machines.
Hospital wants to test by doing a side-by-side comparison: record the proportion of (known) tumors detected by each machine

Let p_0 the proportion found by the old machine
p_1 the " new machine.

a) $H_0: p_0 = p_1$
$H_a: p_0 < p_1$.

b) a Type I error occurs when we decide the new machine is better when it is not.

Real world consequence: spend \$ on new tech that does not help your patient.

c) a Type II error occurs when we decide the ~~old~~ new machine is not better when it is

Real world: we don't purchase machines that can actually help save and prolong lives.

§ 10.3 Z-tests (large sample)

ex: (10.18)
The hourly wages in a particular industry are distributed $N(13.20, 2.50)$.
A company in this industry employs 40 workers, paying them an average of 12.20 \$ per hour.
Can this company be accused of paying in substandard wages? Use $\alpha = 0.01$.

Recall $n = 40 > 30$ is considered "large".
Hence $\bar{x} \sim N(\mu, \sigma^2/n) = N(13.20, 2.50/40)$
$= N(13.20, 0.0625)$.

Our test: $H_0: \mu_c = 13.20$, μ_c the company average
$H_a: \mu_c < 13.20$

Here $\alpha = 0.01$ is the decision rule
~ equivalently, the significance level
~ also the probability of making a Type I error.

$\alpha = P(\text{Type I error})$
$= P(\text{rejecting } H_0 \text{ when it is true})$
$= 0.01$

Compute the p-value.

$p = P(\bar{X} \leq 12.20 \mid \mu = 13.20)$

convert to $Z \sim N(0,1)$ as in 325

$= P\left(\cancel{\bar{X}}\, Z \leq \dfrac{12.20 - 13.20}{0.25} \right)$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \qquad \sigma_{\bar{X}} = \sqrt{\frac{2.50}{40}} = \frac{1}{4}.$$

$= P(Z \leq -4)$

$= P(Z \geq 4) = 0.0000317$ by Table 4

$=$ pnorm(12.20, 13.20, 0.25)

Decision + Conclusion:

p-value $<< \alpha = 0.05$.

reject H_0.

Yes, the company appears to be systemically underpaying its employees in relation to the rest of the industry.

Rmks: ① $Z = \dfrac{\bar{X} - \mu}{\sigma_{\bar{X}}}$ called the <u>test statistic</u>.

② R command pnorm(c, μ, σ)

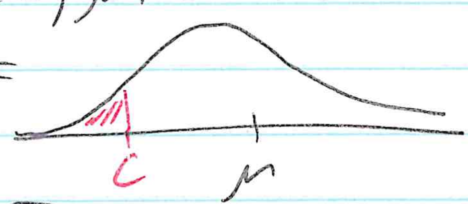
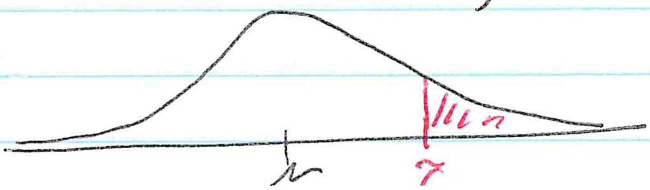
$=$ 

Table 4 : 

ex (10.21) shear strength measurement are derived from unconfined compression tests for two types of soils.

Soil I	Soil II	
$n_1 = 30$	$n_2 = 35$	
$\bar{Y}_1 = 1.65$	$\bar{Y}_2 = 1.43$	(tons per sq. ft)
$s_1 = 0.26$	$s_2 = 0.22$	

Do the soils appear to differ w/ respect to average shear strength at the 1% significance level?

Note $n_1, n_2 \geq 30 \Rightarrow$ we can use $\sigma_1 = s_1$ and $\sigma_2 = s_2$ without loss of precision.

$H_0: \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$

$H_a: \mu_1 \neq \mu_2$ ← this is called a 2-sided test.

In Ch8, saw $\bar{Y}_1 - \bar{Y}_2$ distributed

$$N\left(\mu_1 - \mu_2, \ \sigma_1^2/n_1 + \sigma_2^2/n_2\right).$$

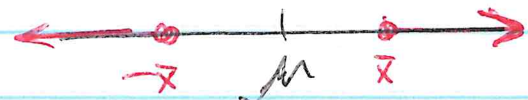
Under the null hypothesis,

$$N\left(0, \frac{(0.26)^2}{30} + \frac{(0.22)^2}{35}\right) = N\left(0, \overset{0.00363}{\cancel{0.0603}}\right)$$

and $\sigma_{\bar{Y}_1 - \bar{Y}_2} = 0.0603$

For a 2-sided test

need prob



$$
\begin{aligned}
p &= P(|\bar{Y}_1 - \bar{Y}_2| > 1.65 - 1.43) \\
&= P(|\bar{Y}_1 - \bar{Y}_2| > 0.22) \\
&= P(\bar{Y}_1 - \bar{Y}_2 < -0.22) + P(\bar{Y}_1 - \bar{Y}_2 > 0.22) \\
&= 2P(\bar{Y}_1 - \bar{Y}_2 > 0.22) \\
&= 2P\left(Z > \frac{0.22}{0.0603}\right) \\
&= 2P(Z > 3.648) \quad \text{not on table 4} \\
&< 2P(Z > 3.5) \\
&= 2 \cdot (0.000233) \\
&= 0.000466 \quad << \alpha = 0.01
\end{aligned}
$$

Conclusion: statistically significant

i.e. supports H_a.

The shear strengths are different.

$$
\begin{aligned}
p &= 2\text{pnorm}(-0.22, 0, 0.0603) \\
&= 0.000265
\end{aligned}
$$

Rmk: In the last 2 examples, using the empirical rule

68-95-99.7.

we could have concluded "reject H_0"

simply on the Z-score alone.

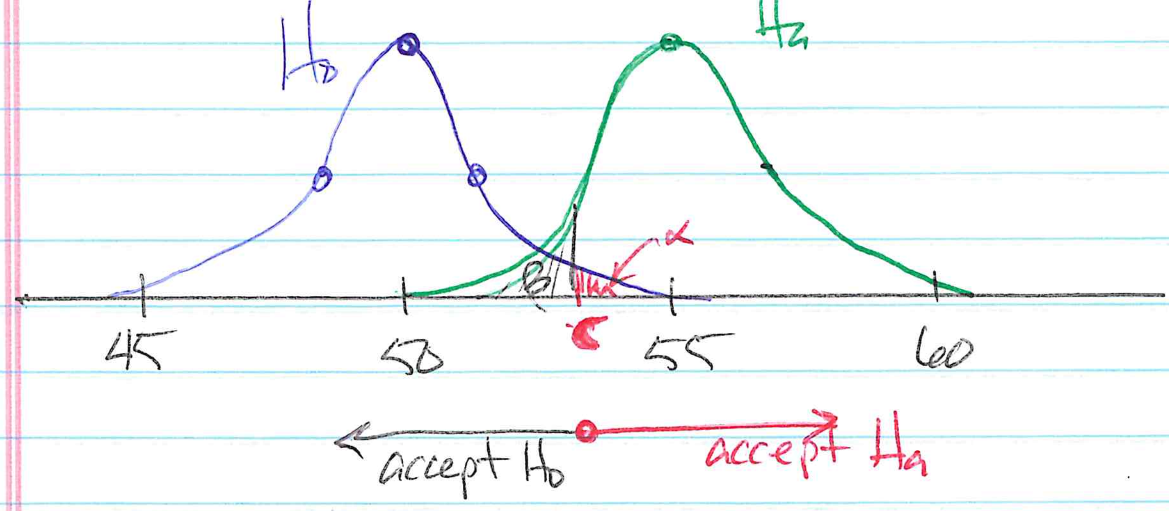
§ 10.4 more about errors + sample size.

moto ex: X equals the breaking strength of a steel bar. If the bar is manufactured by Process I, it is known $\cancel{X} \sim N(50, 36)$.
A new process II is used and it is hoped that the steel is 10% stronger.
That is $X \sim N(55, 36)$

Our test? $H_0: \mu_I = 50$
$H_a: \mu_{II} = 55$

Okay... we can't really test this.
But we can construct a test where if H_a is true, we can minimize (or control) both the Type I and Type II errors.

For the sake of concreteness, lets set $n=16$
Then $\sigma_{\bar{x}}^2 = \frac{36}{16}$ and $\sigma_{\bar{x}} = \frac{6}{4} = 1.5$



$H_0: \mu = 50$
$H_a: \mu > 50$

α = Prob(Type I error)
= Prob(rejecting H_0 when it is true).

On the other hand, given this α, we can also see
β = Prob(Type II error)
= Prob(accept H_0 when H_a is true).

In otherwords, $\alpha = \text{Prob}(\bar{X} > C \mid H_0)$
$= \text{Prob}\left(\frac{\bar{X}-50}{1.5} > \frac{C-50}{1.5}\right)$

and $\beta = \text{Prob}(\bar{X} < C \mid H_a)$
$= \text{prob}\left(\frac{\bar{X}-55}{1.5} < \frac{C-55}{1.5}\right)$

For fixed $n=16$, usually chose α small.
$\alpha = P(\bar{X}-50 > 2\sigma_{\bar{x}}) = 0.025$
Then $C = 50 + 2(1.5) = 53$.
and
$\beta = P(\bar{X} < 53 \mid H_a)$
$= P\left(\frac{\bar{X}-55}{1.5} < -1.33\right)$
$= 0.0918$

Note almost ~~twice~~ four as likely to make a Type II error than a Type I.

Of course decreasing α increases β.

$$\alpha = 0.01 \implies z_{score} = 2.33 \qquad (z_{0.98})$$

$$C = 50 + 2.33(1.5)$$
$$= 53.495$$

$$\beta = \text{Prob}(\bar{X} < 53.495 \mid H_a)$$
$$= \text{Prob}(Z < -1.003)$$
$$= 15.87\,\%$$

Again, note the only way to decrease both α and β is to crank up n.

disc: choosing sample size n.

We consider the 1-sided test $H_0: \mu = \mu_0$
$H_a: \mu > \mu_0$

Fix α at the start.

$$\alpha = P(\bar{Y} > c \text{ when } \mu = \mu_0)$$
$$= P\left(Z > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right)$$
$$= P(Z > z_\alpha), \quad z_\alpha = \frac{c - \mu_0}{\sigma/\sqrt{n}}$$

But then, $\beta = P(\bar{Y} < c \text{ when } \underbrace{\mu = \mu_a}_{})$

as w/ the power fcn, we will need to choose a μ_a.

$$= P\Big(Z < \underbrace{\frac{c-\mu_a}{\sigma/\sqrt{n}}}_{\text{note, this is -ive}}\Big)$$

$$= P(Z < -z_\beta) \quad , \quad -z_\beta = \frac{c-\mu_a}{\sigma/\sqrt{n}}.$$

$$\Longrightarrow c = \mu_0 + z_\alpha \sigma/\sqrt{n} \quad \text{and} \quad c = \mu_a - z_\beta \sigma/\sqrt{n}$$

$$\mu_0 + z_\alpha \sigma/\sqrt{n} = \mu_a - z_\beta \sigma/\sqrt{n}.$$

Solve for $n = \left(\frac{(z_\alpha - z_\beta)\sigma}{\mu_a - \mu_0}\right)^2 = \frac{(z_\alpha - z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}$

Rmk: ① of course all of this has the fudge factor that we don't really know μ_a

② If we did $H_0: \mu_0 = \mu_a$, $H_a: \mu_0 > \mu_a$, we'd get the same n.

"Sample size estimator for a one-sided α-level test" ex: back to steel example.

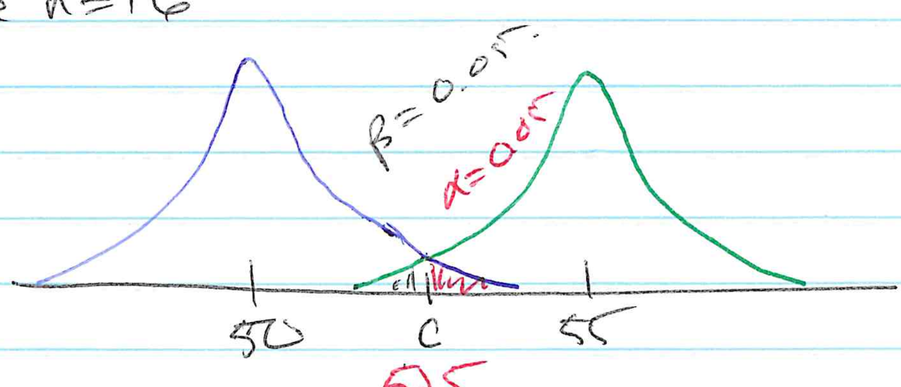
If we decide at the start we want $\alpha = \beta = 0.05$, what n should we choose?

For $\alpha = 0.05 \Rightarrow z_\alpha = 1.645$
Similarly for β, need $z_\beta = 1.645$.

$$n = \left(\frac{2 \cdot 1.645 \cdot 6}{55-50}\right)^2 = 15.5867$$

$= 15.5867$

Choose $n = 16$



§ 10.8 T-tests.

Recall for small sample sizes, we need to use the t-distribution.

ex: 100 ml sample of water from public swimming areas are tested for fecal coliform bacteria. Considered unsafe if the level of bacteria is above 400.

20 samples are taken: $\bar{x} = 1231$
$S = 1038$

Our test: $H_0: \mu_0 = 400$
$H_a: \mu > 400$

The test statistic $t = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}} = \dfrac{1231 - 400}{1038/\sqrt{20}}$
$= 3.580$

The p-value $p = \text{Prob}(T > 3.580)$
not on Table 5.

Recall $df = 20 - 1 = 19$ ...
largest on Table 5 is $t_{0.005} = 2.861$

So $p < \text{Prob}(T > 2.861) = 0.005$

Conclusion: Don't go swimming!

Rmk: In general, at α-level significance

If $H_a := \begin{cases} \mu > \mu_0 \\ \mu < \mu_0 \\ \mu \neq \mu_0 \end{cases} \Longrightarrow RR := \begin{cases} t > t_\alpha \\ t < -t_\alpha \\ |t| > t_{\alpha/2}. \end{cases}$

rejection region

the t-stat.

ex: Lifestyle comparison.
Study monitors the active time (in minutes per day) btwn 2 populations.

		Standing/Walking	S
(obese cohort)	$n=10$	373.269	67.498
(lean cohort)	$n=13$	525.751	107.121

Q: Are these two groups statistically significant?
(i.e. is there evidence of different behavior?)

Our test $H_0: \mu_O = \mu_L$ $\quad \mu_O$, obese, μ_L, lean

$H_a: \mu_O < \mu_L$ ← assumed one-sided...
lean group more active.

But we need to recast this as a difference of means.

$$H_0: \mu_O - \mu_L = 0$$
$$H_a: \mu_O - \mu_L < 0$$

10 pts.

Recall the $100(1-\alpha)\%$ C.I for this is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $S_p = \sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$

w/ $df = n_1 + n_2 - 2$.

Thus our test statistic is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

$\bar{X}_0 - \bar{X}_L = -152.482$.

$$S_p^2 = \frac{(n_0-1)S_0^2 + (n_L-1)S_L^2}{n_0+n_L-2}$$

$$= \frac{(10-1)\,67.498^2 + (13-1)(107.121)^2}{10+13-2}$$

$= 8509.65$

$S_p = 92.2478$

$$T = \frac{-152.482 - 0}{92.2478\sqrt{\frac{1}{10}+\frac{1}{13}}} = -3.9237$$

For $P(T < -3.9237)$
$= P(T > 3.9237)$
$< P(T > 3.768)$
$= 0.0005$ by Table.

super small p-value $\Rightarrow$ reject null hypothesis.
$\mu_0 < \mu_L$.

aside: Cautions on using t-tests.

① To use t-tests, the population needs to at least 20 times the sample size

② If $n < 40$, the original population must normal (or nearly normally) distributed.

If $n > 40$, good to go with any original population distribution.

§ 10.10 The Neyman-Pearson lemma

disc: vocab refresh and improvements.

In §10.4 we considered the simple hypothesis test

$$H_0: \theta = \theta_0$$
$$H_a: \theta = \theta_a \quad (\neq \theta_0)$$

and used assumptions to use sample size to control Type II error.
In real-life, often can't control sample size.

We revisit the power fcn.

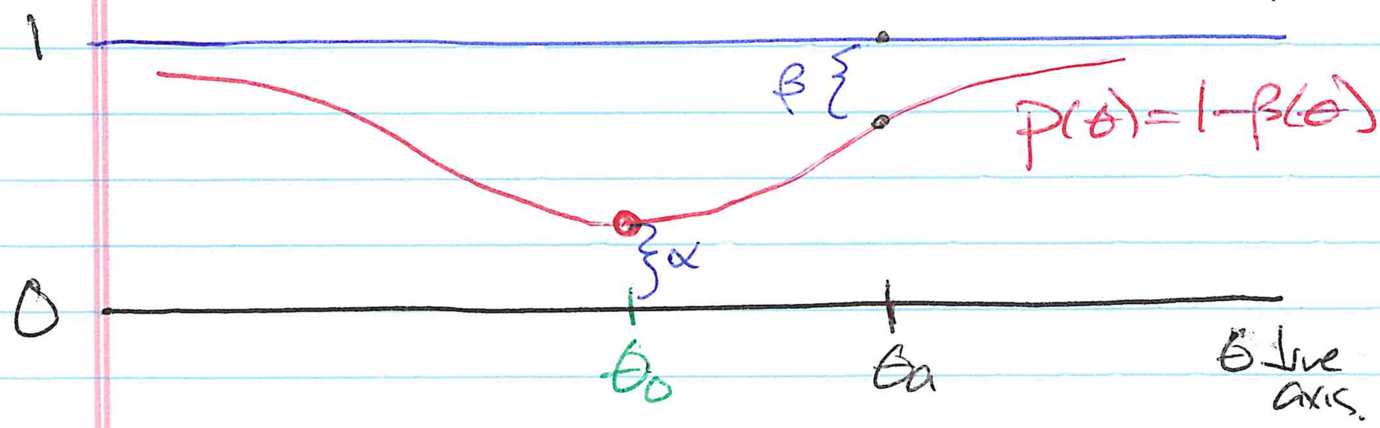
Recall $\alpha = \text{Prob}(\text{reject } H_0 \mid H_0 \text{ true})$ Type I
$\beta = \text{Prob}(\text{accept } H_0 \mid H_a \text{ true})$ Type II

We also talked about the power of a test:

$$\text{power}(\theta) = 1 - \beta(\theta)$$
$$= \text{Prob}(\text{accept } H_a \mid H_a \text{ true})$$
$$= \text{Prob}(\text{accept } H_a \mid \theta = \theta_a)$$

Recall we saw in a first day example that the closer θ_a is to our assumption θ_0, the larger $\beta(\theta_a)$ is and the smaller the power.
For the test $\theta_0 \neq \theta_a$, the graph of ~~of~~ the power fcn has a common form

10pdd



In designing a test, we would like power(θ) to be maximized ... i.e. least likely to make a Type II error.

Some brainy mugs figured this out

topic: Neyman - Pearson Lemma

The last goal of this chapter is to learn how to design the best possible test for simple hypotheses.

Note: $\mu = \mu_0$ is simple
$\mu > \mu_0$ is composite.

To do this "right" we need a little dimensional slight-of-hand.

Let $X_1, \dots, X_n$ be an iid random sample, $X_i \sim f(x_i|\theta)$
Consider the space $X^n : (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$.

def: Given a simple test $H_0: \theta = \theta_0$, $H_a: \theta = \theta_a$, define $C \subseteq X^n \subseteq \mathbb{R}^n$ to be a critical region of size α if

$$\alpha = P(C|\theta_0)$$
$$= \int_C \underbrace{\prod f(x_i|\theta_0)}\, dx_1 \dots dx_n.$$

Rmk: Hey! That is a likelihood fn.

Moreover, a critical region C is called a best critical region of size α if for every other critical region D of size α, we have that

$$P(C|\theta_a) \geq P(D \geq \theta_a)$$

That is, when $H_a: \theta = \theta_a$ is true, the probability of rejecting $H_0: \theta = \theta_0$ using critical region C as the rejection region is at least as great as the corresponding prob. using any other critical region

Thm: The Neyman-Pearson lemma.

let $X_1, \dots, X_n$ be iid w/ pdf $f(x|\theta)$ where θ_0 and θ_a are two possible values of θ. If there exists a positive constant k and a subset C of the sample space such that

① $P\{(X_1, X_2, \dots, X_n) \in C | \theta_0\} = \alpha$

② $\dfrac{L(\theta_0)}{L(\theta_a)} \le k$ for $(x_1, \dots, x_n) \in C$

③ $\dfrac{L(\theta_0)}{L(\theta_a)} \ge k$ for $(x_1, \dots, x_n) \in \overline{C}$

then C is a best critical region of size α for testing simple null hypothesis $H_0: \theta = \theta_0$ against simple alternative hypothesis $H_a: \theta = \theta_a$.

* before proof... example.

ex: Let $Y_1, Y_2, \dots, Y_n$ from $f(y|\theta) = \frac{2}{\theta} y e^{-y^2/\theta}$, $y > 0$

(the Rayleigh dist'n)

Want to test $H_0: \theta = \theta_0$

$H_a: \theta = \theta_a$.

(WLOG, we assume $\theta_a > \theta_0$)

Note $L(y_1, \dots, y_n | \theta) = \underbrace{\left(\frac{2}{\theta}\right)^n \exp\left(-\sum_1^n y_i^2/\theta\right)}_{g} \cdot \underbrace{\prod y_i}_{h}$

future flash:

10 pts

The N-P ratio of likelihood becomes.

$$\frac{L(\theta_0)}{L(\theta_a)} = \frac{\left(\frac{2}{\theta_0}\right)^n \exp\left(-\sum y_i^2/\theta_0\right) \cdot \prod y_i}{\left(\frac{2}{\theta_a}\right)^n \exp\left(-\sum y_i^2/\theta_a\right) \cdot \prod y_i}$$

$$= \left(\frac{\theta_a}{\theta_0}\right)^n \exp\left(-\sum y_i^2 \left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)\right)$$

We want C to relate to our rejection region RR.
So ② ⇒ " reject H_0 if

$$\left(\frac{\theta_a}{\theta_0}\right)^n \exp\left[-\sum y_i^2 \left(\frac{1}{\theta_0} - \frac{1}{\theta_a}\right)\right] < k \text{ "}.$$

This looks scary as hell... but θ_a, θ_0 are fixed and since $\theta_0 < \theta_a$, $\frac{1}{\theta_0} - \frac{1}{\theta_a} > 0$.

So making $\frac{L(\theta_0)}{L(\theta_a)}$ "small" requires $\sum y_i^2$ to be large enough.

i.e. we need $\sum_1^n y_i^2 > k'$ for some k'.

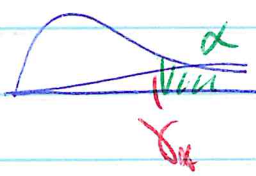
New Problem! To determine k', we need to know how $S = \sum_1^n y_i^2$ is distributed

Using the CDF method (§6.3) we can show that the distribution of Y_i^2 is exponential w/ mean θ and thus $\sum Y_i^2 \sim$ Gamma$(n, 1/\theta)$

Stop! What just happened?

The N-P lemma tells us that the most powerful test of H_0 vs. H_a is to use the statistic $S = \sum Y_i^2$

Given any α level significance, we use the test ~~test~~



S is larger than the $100(1-\alpha)$ percentile of the Gamma distribution Gamma$(n, 1/\theta_0)$ distribution.

S larger ... reject the null hypothesis.

Moreover, by the theorem, we don't have to compare the power of other possible tests, because the lemma says any data $(Y_1, \ldots, Y_n) \in C$ defined by our α significance level $\leftrightarrow$ $100(1-\alpha)$ percentile of Gamma$(n, 1/\theta_0)$ results in the most powerful test

$$P(C \mid \theta_a) \geq P(D \mid \theta_a)$$

def: The test using the best critical region is called a <u>most powerful test</u>

Recap: We wanted to construct a hypothesis test at α-significance level.

All in one swoop, the P-N lemma says

1. here is the stat to use
2. here is your rejection region
3. this is in fact the most powerful test possible.

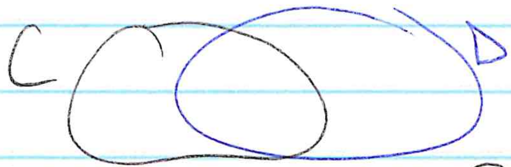
(This is kinda awesome.)

Thm: NP Lemma.

Pf: (continuous case) $B \subseteq \mathbb{R}^n$ notation:
Let $\int_B L(\theta) = \int \cdots \int_B L(x_1, x_2, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n$.
Assume $\exists$ C satisfying 1, 2 + 3.
Then given fixed α, $\alpha \overset{!}{=} \int_C L(\theta_0)$,

Now assume D is another critical region, $\alpha = \int_D L(\theta_0)$

So $0 = \int_C L(\theta_0) - \int_D L(\theta_0)$



$$0 = \int_C L(\theta_0) - \int_D L(\theta_0)$$

$$= \int_{C\cap\bar{D}} L(\theta_0) + \int_{C\cap D} L(\theta_0)$$

$$- \int_{C\cap D} L(\theta_0) - \int_{\bar{C}\cap D} L(\theta_0)$$

$$= \int_{C\cap\bar{D}} L(\theta_0) - \int_{\bar{C}\cap D} L(\theta_0)$$

By hypo ② $\exists$ k s.t. $kL(\theta_a) \geq L(\theta_0)$ at each point in C.

So $k\int_{C\cap\bar{D}} L(\theta_a) \geq \int_{C\cap\bar{D}} L(\theta_0)$

Also by ③, we have $L(\theta_0) \geq k L(\theta_a)$ at each point in $\bar{C}$.

So $k\int_{\bar{C}\cap D} L(\theta_a) \leq \int_{\bar{C}\cap D} L(\theta_0)$

So $0 = \int_{C\cap\bar{D}} L(\theta_0) - \int_{\bar{C}\cap D} L(\theta_0)$

$$\leq k\int_{C\cap\bar{D}} L(\theta_a) - k\int_{\bar{C}\cap D} L(\theta_a)$$

$$= k\Big[\int_{C\cap\bar{D}} L(\theta_a) + \int_{C\cap D} L(\theta_a)$$

$$- \int_{C\cap D} L(\theta_a) - \int_{\bar{C}\cap D} L(\theta_a)\Big]$$

10 pg 29.

$$= k\left[\int_C L(\theta_a) - \int_D L(\theta_a)\right]$$

$$\Rightarrow \int_C L(\theta_a) \geq \int_D L(\theta_a).$$

i.e. $P(C \mid \theta_a) \geq P(D \mid \theta_a)$ for all critical regions D of α.

Hence C is a best critical region of size α. ∎

ex: Back to our old sample size example.

$X_1, \ldots, X_n \sim N(\mu, 36)$

We played w/ $H_0: \mu = 50$ vs $H_a: \mu = 55$

Consider the ratio of likelihood fns.

$$\frac{L(50)}{L(55)} = \frac{(72\pi)^{-n/2} \exp\left[-\left(\frac{1}{72}\right)\sum_1^n (x_i - 50)^2\right]}{(72\pi)^{-n/2} \exp\left[-\left(\frac{1}{72}\right)\sum_1^n (x_i - 55)^2\right]}$$

$$= \exp\left[-\frac{1}{72}\sum_1^n \underbrace{\left[(x_i-50)^2 - (x_i-55)^2\right]}_{10x_i - 525}\right]$$

$$= \exp\left[\frac{-1}{72}\sum_1^n (10x_i - 525)\right]$$

$$= \exp\left[-\frac{5}{36}\sum_1^n x_i + \frac{175n}{24}\right] \le k.$$

$$\Rightarrow -\frac{5}{36}\sum_1^n x_i + \frac{175n}{24} \le \ln k.$$

$$\Rightarrow \sum_1^n x_i \ge \frac{105}{2}n - \frac{36}{5}\ln k$$

So $\bar{X} = \frac{1}{n}\sum_1^n x_i \ge \frac{105}{2} - \frac{36}{5n}\ln k.$

We have our stat!

Equivalently, we have $\bar{X} \ge k' = \frac{105}{2} - \frac{36 \ln k}{5n}$.

This results in designing a best critical region, by $C = \{(x_1, \ldots x_n) : \bar{X} \ge k'\}$.

where k' is selected so that the size of the ~~critical~~ rejection region is α.

e.g. Say $n = 16$ and $k' = 53$ (like before).

Since $\bar{X} \sim N(50, 36/16)$, under H_0, we have

$$\alpha = P(\bar{X} \ge 53 \mid \mu = 50)$$

$$= P\left(\frac{\bar{X} - 50}{6/4} \ge \frac{3}{6/4} \,\Big|\, \mu = 50\right)$$

$$= P(Z \ge 2 \mid \mu = 50)$$

$$= 1 - P(Z \le 2) = 0.0228$$

If we were to choose RR to be $\bar{X} \geq 53$, then $\alpha = 0.0228$ and using the stat $\bar{X}$ is the best test for $H_0: \mu = 50$ vs $H_a: \mu = 55$.

Rmk: This example demonstrates something that is often true $\frac{L(\Theta_0)}{L(\Theta_a)} \leq k$ can be expressed in terms of a stat $u(x_1, \ldots, x_n)$ s.t. $u(x_1, \ldots, x_n) \leq k'$ or $\geq k'$, where k' is selected so that the critical region ~~α~~ size α.
i.e. the entire test is based on the statistic u.

eg. $\alpha = 0.05$ then we want

$$0.05 = P(\bar{X} \geq k' : \mu = 50)$$
$$= P\left(\frac{\bar{X} - 50}{6/4} \geq \frac{k' - 50}{6/4}\right)$$
$$= 1 - P\left(Z \leq \frac{k' - 50}{6/4}\right)$$
$$\Rightarrow 1.645 = \frac{k' - 50}{6/4} \Rightarrow k' = 52.47$$

ex: do it again except let $H_a: \mu_a > \mu_0 = 50$
a composite hypo.

Need $\frac{L(50)}{L(\mu_a)} \le k$.

$$\frac{L(50)}{L(\mu_a)} = \exp\left[\frac{-1}{72}\left(\sum(x_i-50)^2 - \sum(x_i-\mu_a)^2\right)\right]$$
$$= \exp\left[\frac{-1}{72}\left(2(\mu_a-50)\sum_1^n x_i + n(50^2-\mu_a^2)\right)\right]$$

Like before $\frac{L(50)}{L(\mu_a)} \le k$ requires $\sum_1^n x_i$ large enough

i.e. $\bar{X} \ge \frac{-72 \ln k}{2n(\mu_a-50)} + \frac{50+\mu_a}{2} = k'$

As before, best critical region of size α for testing
$H_0: \mu_0 = 50$
vs $H_a: \mu_a > 50$
is given by $C = \{(x_1, \dots, x_n) : \bar{x} > k'\}$
where k' is selected such that
$$P(\bar{X} \ge k' \mid H_0: \mu = 50) = \alpha.$$

Note the same value of k' can be used for each $\mu_a > 50$ (of course original k does not remain the same).
(e.g. $\alpha = 0.05 \Rightarrow k' = 52.47$ again)

Since the critical region C defines a test that is most powerful against each simple alternative $\mu_a > \mu_0$, this is called a

<u>uniformly most powerful test</u>.

def. The test using the best critical region is called a <u>most powerful test</u>

A test defined by a critical region C of size α, is a <u>uniformly most powerful test</u> (UMPT). if it is a most powerful test against each simple alternative in H_a.

The critical region C is called a <u>uniformly most powerful critical region of size</u> α.

The Point ~~Rmk~~: Our Z-test and T-tests are UMPT.