Statistical Inference

Statistical inference provides us with the process of testing "hypotheses" under investigation from data. Procedures of statistical inference are given for (i) parameter estimates and their confidence intervals, and (ii) tests of statistical hypotheses. Some common tests, called t-test, z-test, and χ^2-test, are associated respectively with t-distribution, normal distribution, and χ^2-distribution.

Population distribution and parameters.

A *random sample*

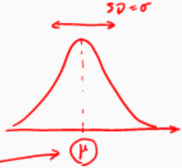
$$\overbrace{X_1, \ldots, X_n}^{\text{DATA = iid random variables}}$$

is regarded as independent and identically distributed (iid) random variables governed by pdf $f(x;\theta)$. A value θ (which is a vector of real numbers in general) represents the characteristics of this underlying *population distribution*, and is called a *parameter*. Suppose, for example, that the underlying distribution is the normal distribution with (μ, σ^2). $= \theta$ Then the values μ and σ^2 are the parameters.

Statistics and point estimates.

A random sample is viewed as a random vector

sample



$$\boldsymbol{X} = (X_1, \ldots, X_n),$$

guess μ

and a random variable $u(\boldsymbol{X})$ constructed from the random vector $\boldsymbol{X}$ is called a *statistic*. For example, the sample mean $\bar{X}$ is a statistic. A *point estimate* is a statistic $u(\boldsymbol{X})$ which is a "best guess" for the true value θ. Suppose that the underlying distribution is the normal distribution with (μ, σ^2). Then the sample mean $\bar{X}$ is in some sense a best guess of the parameter μ.

Risk function and bias.

$\mathbb{X} = (X_1, \cdots, X_n) \xrightarrow{u} u(\mathbb{X})$

Let $u(\boldsymbol{X})$ be a point estimate for θ. Then the functional $R(\theta, u) = E[(u(\boldsymbol{X}) - \theta)^2]$ of u is called the *mean square-error risk function*. We can immediately observe that

$$R(\theta, u) \overset{?}{=} \underbrace{\mathrm{Var}(u(\boldsymbol{X}))}_{\text{minimize it}} + [\underbrace{E(u(\boldsymbol{X})) - \theta}_{\text{bias} \leftarrow \text{make it zero} \leftarrow \text{1st step}}]^2 = \mathrm{Var}(u(\boldsymbol{X})) + [b(\theta, u)]^2,$$

where $b(\theta, u) = E(u(\boldsymbol{X})) - \theta$ is called the *bias* of $u(\boldsymbol{X})$. One of the important attributes of point estimate is unbiasedness. Since a statistic $u(\boldsymbol{X})$ is a random variable, we can consider the expectation $E[u(\boldsymbol{X})]$. Then the point estimate $u(\boldsymbol{X})$ of θ is called an *unbiased estimator* if it satisfies $E[u(\boldsymbol{X})] = \theta$. For example, the sample mean $\bar{X}$ is an unbiased estimate of the mean μ, since $E(\bar{X}) = \mu$.

$\{A + B\}^2 = A^2 + 2AB + B^2$

$$R(\theta, u) = E[(u(\mathbb{X}) - \theta)^2] = E[\{(u(\mathbb{X}) - E[u(\mathbb{X})]) + (\underbrace{E[u(\mathbb{X})] - \theta}_{\text{constant}})\}^2]$$

$$= \underbrace{E[(u(\mathbb{X}) - E[u(\mathbb{X})])^2]}_{\mathrm{Var}(u(\mathbb{X}))} + 2\underbrace{E[(u(\mathbb{X}) - E[u(\mathbb{X})])]}_{0}(\underbrace{E[u(\mathbb{X})] - \theta}_{\text{constant}}) + (E[u(\mathbb{X})] - \theta)^2$$

Confidence interval.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from $f(\boldsymbol{x}; \theta)$. Let $u_1(\boldsymbol{X})$ and $u_2(\boldsymbol{X})$ be statistics satisfying $u_1(\boldsymbol{X}) \leq u_2(\boldsymbol{X})$. If

$$P\left(u_1(\boldsymbol{X}) < \theta < u_2(\boldsymbol{X})\right) = 1 - \alpha \qquad \text{for every } \theta,$$

then the random interval $(u_1(\boldsymbol{X}), u_2(\boldsymbol{X}))$ is called a *confidence interval of level* $(1 - \alpha)$. Choice of α is either $\alpha = 0.1$, 0.05, 0.01

Population mean under normal assumption.

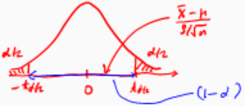
Let $X_1, \dots, X_n$ be a random sample from $N(\mu, \sigma)$. The sample mean $\bar{X}$ is an unbiased estimate of the parameter μ. Then the random variable $\dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ has the t-distribution with $(n-1)$ degrees of freedom. Thus, by using the critical point $t_{\alpha/2, n-1}$

$$P\left(\left|\frac{\bar{X} - \mu}{S/\sqrt{n}}\right| < t_{\alpha/2,n-1}\right) = P\left(\bar{X} - \frac{t_{\alpha/2,n-1} S}{\sqrt{n}} < \mu < \bar{X} + \frac{t_{\alpha/2,n-1} S}{\sqrt{n}}\right) = 1 - \alpha$$

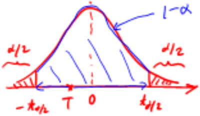
has the probability of $(1 - \alpha)$. This implies that the parameter μ is in the interval

$$\left(\bar{X} - \frac{t_{\alpha/2,n-1} S}{\sqrt{n}},\ \bar{X} + \frac{t_{\alpha/2,n-1} S}{\sqrt{n}}\right)$$

with probability $(1 - \alpha)$. The interval is also known as the t-interval.



$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t\text{-distribution}$$



$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

$$\Updownarrow \qquad\qquad \Updownarrow$$

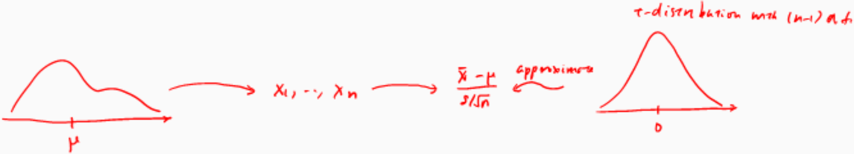
$$\mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}} \qquad \bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu$$

$$P\left(\mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}} \text{ and } \bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu\right) = 1-\alpha$$

$$\Updownarrow$$

$$P\left(\underbrace{\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}}_{u_1(X)} < \mu < \underbrace{\bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}}_{u_2(X)}\right) = 1-\alpha$$

Normal assumption is not necessary.



Even if a random sample $X_1, \dots, X_n$ is not normally distributed, the central limit theorem says that the estimate $\bar{X}$ is approximately distributed as $N(\mu, \sigma^2/n)$ when n is large. In either case it is sensible to construct a confidence interval with critical point $t_{\alpha/2, n-1}$ from t-distribution.

Usually n is larger than 30.

Example

A random sample of n milk containers is selected, and their milk contents are weighed. The data

$$X_1, \ldots, X_n \tag{8.1}$$

can be used to investigate the unknown population mean of the milk container weights. A random sample can be assumed to be iid normal distribution. Suppose that we have calculated $\bar{X} = 2.073$ and $S = 0.071$ from the actual data with $n = 30$. Then construct 95% confidence interval.

$(1-\alpha)\%$

$\Uparrow$

$\alpha = 0.05$

Example

A random sample of n milk containers is selected, and their milk contents are weighed. The data

$$X_1, \ldots, X_n \tag{8.1}$$

can be used to investigate the unknown population mean of the milk container weights. A random sample can be assumed to be iid normal distribution. Suppose that we have calculated $\bar{X} = 2.073$ and $S = 0.071$ from the actual data with $n = 30$. Then construct 95% confidence interval.

By choosing $\alpha = 0.05$, we have the critical point $t_{\underbrace{0.025}_{\alpha/2},29} = 2.045$, and therefore, obtain the confidence interval

$$\left(2.073 - \frac{2.045 \times 0.071}{\sqrt{30}},\ 2.073 + \frac{2.045 \times 0.071}{\sqrt{30}}\right) = (2.046,\ 2.100)$$

of level 0.95 (or, of level 95%).

Population proportion.

$X_i = \begin{cases} 1 & \text{with } p \\ 0 & \text{with } (1-p) \end{cases} \longrightarrow Y = X_1 + \cdots + X_n = \#\text{ of successes} \sim \text{Binomial with } (n, p) \overset{\text{CLT}}{\sim} N(\mu, \sigma^2)$

$\overset{\text{approx}}{\sim} N(np, np(1-p))$

$E(Y)$ $Var(Y)$

Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with success probability p. The sample mean $\bar{X}$ is an estimate of the parameter p, and unbiased. By the central limit theorem, the random variable

$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{Y}{n} \overset{\text{approx}}{\sim} N\left(\overset{\mu}{p}, \overset{\sigma^2}{\frac{p(1-p)}{n}}\right)$

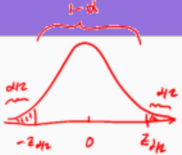
$E(\bar{X})$ $Var(\bar{X})$

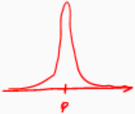
$$\left(\frac{X-\mu}{\sigma} = Z = \right)\frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$$

$Z = \frac{\bar{X} - \mu}{\sigma}$

has approximately $N(0, 1)$ as n gets larger (at least $np > 5$ and $n(1-p) > 5$ by rule of thumb).

Critical point. Here we define the critical point z_α for standard normal distribution by $P(X > z_\alpha) = \alpha$ with standard normal random variable X.





$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$

$$1-\alpha = P\left(\left|\frac{\bar{X}-p}{\sqrt{p(1-p)/n}}\right| < z_{\alpha/2}\right)$$
$$= P\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \bar{X} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

has the approximate probability of $1-\alpha$. Here we can use $\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$ as an estimate for $\sqrt{\frac{p(1-p)}{n}}$. Together we obtain the confidence interval

$$\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \; \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right)$$

of level $(1-\alpha)$.

Alternative confidence interval.

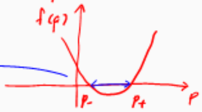
There is an alternative and possibly more accurate method to derive a confidence interval. Here we observe that

$$1-\alpha \simeq P\left(\left|\frac{\bar{X}-p}{\sqrt{p(1-p)/n}}\right| < z_{\alpha/2}\right) = P\left(\frac{(\bar{X}-p)^2}{p(1-p)/n} < z_{\alpha/2}^2\right) = P\left((\bar{X}-p)^2 < z_{\alpha/2}^2\, p(1-p)/n\right)$$

$$P(p_- < p < p_+) = P\left(\underbrace{\left(n+z_{\alpha/2}^2\right)p^2 - 2\left(n\bar{X}+z_{\alpha/2}^2/2\right)p + n\bar{X}^2}_{} < 0\right) \approx 1-\alpha.$$

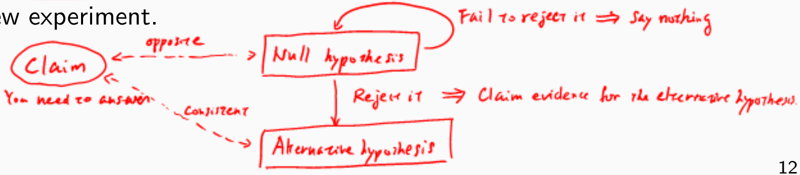
This implies that the parameter p is in the interval $(\hat{p}_-, \hat{p}_+)$ with probability $(1-\alpha)$, where

$$\hat{p}_\pm = \frac{n\bar{X} + z_{\alpha/2}^2/2 \pm z_{\alpha/2}\sqrt{n\bar{X}(1-\bar{X}) + z_{\alpha/2}^2/4}}{n+z_{\alpha/2}^2}.$$



Concept of statistical hypotheses.

Suppose that a researcher is interested in whether the new drug works. The process of determining whether the outcome of the experiment points to "yes" or "no" is called *hypothesis testing*. A widely used formalization of this process is due to Neyman and Pearson. Our hypothesis is then the *null hypothesis* that the new drug has no effect —the null hypothesis is often the reverse of what we actually believe, why? Because the researcher hopes to reject the hypothesis and announce that the new drug leads to *significant* improvements. If the hypothesis is *not* rejected, the researcher announces *nothing* and goes on to a new experiment.



Hypothesis test for population mean.

Hospital workers are subject to a radiation exposure emanating from the skin of the patient. A researcher is interested in the plausibility of the statement that the population mean μ of radiation level is μ_0 $= 5.4$ —the researcher's hypothesis. Then the *null hypothesis* is

$$H_0 : \ \mu = \mu_0. \quad \xleftrightarrow{\text{opposite}} \text{Claim}$$

The "opposite" of the null hypothesis, called an *alternative hypothesis*, becomes

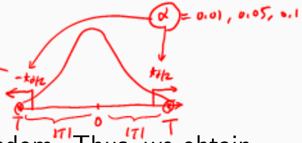
$$H_A : \ \mu \neq \mu_0. \quad \xleftrightarrow{\text{consistent}} \text{Claim}$$

Thus, the hypothesis testing problem "H_0 versus H_A" is formed. The problem here is to whether or not to reject "H_0 in favor of H_A."

To assess this hypothesis, the radiation levels $X_1, \ldots, X_n$ are measured from n patients who had been injected with a radioactive tracer, and assumed to be independent and normally distributed with the mean μ. Under the null hypothesis, the random variable

$X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$



has the t-distribution with $(n-1)$ degrees of freedom. Thus, we obtain the exact probability

This is unlikely under H_0 ⟶ Reject H_0

$$P\left(|T| \geq t_{\alpha/2,n-1}\right) = \alpha.$$

When α is chosen to be a small value (0.05 or 0.01, for example), it is *unlikely* that the absolute value $|T|$ is larger than the critical point $t_{\alpha/2,n-1}$. Then we say that the null hypothesis H_0 is *rejected* with *significance level* α (or, *size* α) when the observed value t of T satisfies $|t| > t_{\alpha/2,n-1}$.

Example

We have $\mu_0 = 5.4$ for the hypothesis, and decided to give a test with significance level $\alpha = 0.05$. Suppose that we have obtained $\bar{X} = 5.145$ and $S = 0.7524$ from the actual data with $n = 28$.

$H_0: \mu = 5.4 \xrightarrow{\text{Reject } H_0} H_A: \mu \neq 5.4$

Example

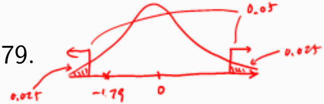
We have $\mu_0 = 5.4$ for the hypothesis, and decided to give a test with significance level $\alpha = 0.05$. Suppose that we have obtained $\bar{X} = 5.145$ and $S = 0.7524$ from the actual data with $n = 28$.

Then we can compute

under H_0

$$T = \frac{5.145 - 5.4}{0.7524/\sqrt{28}} \approx -1.79.$$

t-distribution



Since $|T| = 1.79 \leq t_{0.025,27} = 2.052$, the null hypothesis cannot be rejected. Thus, the evidence against the null hypothesis is not persuasive.

Test statistic and p-value.

The above random variable T is called the *t-statistic*, or "*test*" statistic. Having observed "$T = t$," we can calculate the *p-value*

$$p^* = P(|Y| \geq |t|) = 2 \times P(Y \geq |t|),$$

where the random variable Y has a t-distribution with $(n-1)$ degrees of freedom. Then we have the relation "$p^* < \alpha \Leftrightarrow |t| > t_{\alpha/2,n-1}$." Thus, we reject H_0 with significance level α when $p^* < \alpha$. In the above example, we can compute the p-value

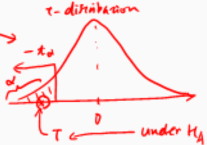
$$p^* = 2 \times P(Y \geq 1.79) \approx 0.0847 \geq 0.05;$$

thus, we cannot reject H_0.

One-sided hypothesis test.

under H_0

$H_0: \mu = \mu_0$ vs $H_A: \mu < \mu_0$

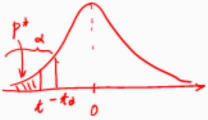


In the same case of hospital workers subject to a radiation exposure, this time the researcher is interested in the plausibility of the statement that the population mean μ is ~~greater~~ less than μ_0. Then the hypothesis testing problem is

$$H_0: \ \mu = \mu_0 \quad \text{versus} \quad H_A: \ \mu < \mu_0.$$

The same t-statistic $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ is used as a *test statistic*. And we reject H_0 with significant level α when you find that $t < -t_{\alpha,n-1}$ for the observed value t of T.

One-sided hypothesis test, continued.



Once T=t is observed

Alternatively we can construct the p-value

$$p^* = P(Y \leq t),$$

where the random variable Y has a t-distribution with $(n-1)$ degrees of freedom. Because of the relation "$\underline{p^* < \alpha} \Leftrightarrow \underline{t < -t_{\alpha,n-1}}$," we can reject H_0 with significant level α when $p^* < \alpha$.

Example

We use the same $\mu_0 = 5.4$ for the hypothesis and the same significance level $\alpha = 0.05$, but use the one-sided test. Recall that $\bar{X} = 5.145$ and $S = 0.7524$ were obtained from the data with $n = 28$.

$H_0: \mu = 5.4$ $\xrightarrow{\text{Reject } H_0}$ $H_A: \mu < 5.4$ $\leftarrow$ $\bar{x} < 5.4$

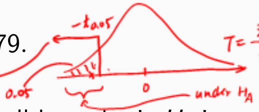
Example

We use the same $\mu_0 = 5.4$ for the hypothesis and the same significance level $\alpha = 0.05$, but use the one-sided test. Recall that $\bar{X} = 5.145$ and $S = 0.7524$ were obtained from the data with $n = 28$.

1. Then we compute

under H_0 → T-distribution

$$T = \frac{5.145 - 5.4}{0.7524/\sqrt{28}} \approx -1.79.$$



$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, $\mu_0 = 5.4$

Since $T = -1.79 < -t_{0.05,27} = -1.703$, the null hypothesis H_0 is rejected. Thus, the outcome is statistically significant so that the population mean μ is smaller than 5.4.

2. Alternatively, we can find the p-value $p^* = P(Y \leq -1.79) \approx 0.0423 < 0.05$; thus, the null hypothesis should be rejected.

One-sided hypothesis test: Opposite case.

under H_0

We can also consider the hypothesis testing problem

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_A: \mu > \mu_0.$$

under H_A $\rightarrow$ $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

1. Using the t-statistics $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$, we can reject H_0 with significant level α when the observed value t of T satisfies $t > t_{\alpha, n-1}$.
2. Alternatively we can construct the p-value $p^* = P(Y \geq t)$ with the random variable Y has a t-distribution with $(n-1)$ degrees of freedom. Because of the relation "$p^* < \alpha \Leftrightarrow t > t_{\alpha,n-1}$," we can reject H_0 when $p^* < \alpha$.

Summary.

When the null hypothesis H_0 is rejected, it is reasonable to find out the confidence interval of the population mean μ. The following table shows the confidence interval we can construct when your null hypothesis is rejected. Here $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ is the test statistic, and α is the significance level of your choice.

It does not contain μ_0

Test	When to reject	$(1-\alpha)$-level confidence interval
$H_A : \mu \neq \mu_0$	$\lvert T \rvert > t_{\alpha/2,n-1}$	$\left(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$
$H_A : \mu > \mu_0$	$T > t_{\alpha,n-1}$	$\left(\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}},\ \infty\right)$
$H_A : \mu < \mu_0$	$T < -t_{\alpha,n-1}$	$\left(-\infty,\ \bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right)$

Type I error: Two-sided test.

We define a function $K(\theta)$ of parameter θ by the probability that H_0 is rejected given $\mu = \theta$.

$H_0 : \mu = \mu_0$

$$K(\theta) = P(\text{"Reject } H_0\text{"} \mid \mu = \theta)$$

is called the *power function*. What is the probability that we incorrectly reject H_0 when it is actually true? Such an error is called *type I error*, when $\mu = \mu_0$ and the probability of type I error is exactly the significant level α, as explained in the following: The probability of type I error for the two-sided hypothesis test is given by $K(\mu_0)$. Then we have

H_0 is True

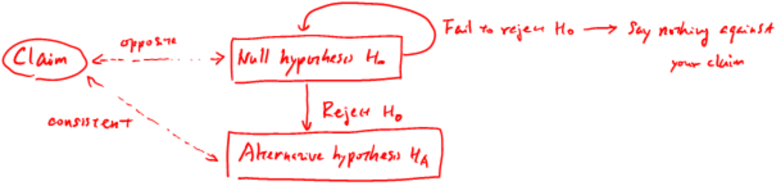
$$K(\mu_0) = P\left(|T| \geq t_{\alpha/2,n-1}\right) = \alpha.$$

		Truth	
		H_0 is true	H_0 is false
Decision	Reject H_0	Type I error	Power of test
	Fail to reject H_0		Type II error

$1-\beta$

α = chance to reject incorrectly

β = chance to fail to reject H_0 when H_0 is false

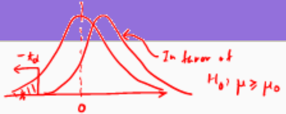


Type I error: One-sided test.

under H_0

$H_0: \mu = \mu_0$ vs. $H_A: \mu < \mu_0$

under H_A



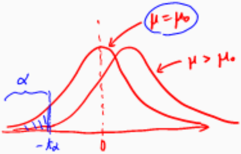
In one-sided hypothesis test, the probability of type I error is the worst (that is, largest possible) probability $\max_{\theta \geq \mu_0} K(\theta)$ of type I error. Given $\mu = \theta$, the random variable

$$\frac{\bar{X} - \theta}{S/\sqrt{n}} = T - \frac{\theta - \mu_0}{S/\sqrt{n}} = T - \delta$$

has the t-distribution with $(n-1)$ degrees of freedom, where $\delta = \dfrac{\theta - \mu_0}{S/\sqrt{n}}$. By observing that $\delta \geq 0$ if $\theta \geq \mu_0$, we obtain

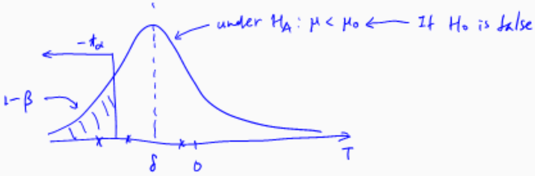
$$K(\theta) = P(T \leq -t_{\alpha,n-1}) \leq P(T - \delta \leq -t_{\alpha,n-1}) = \alpha.$$

Thus, we obtain $\max_{\theta \geq \mu_0} K(\theta) = \alpha$.



Power of test.

What is the probability that we incorrectly accept H_0 when it is actually false? Such probability β is called the probability of *type II error*. Then the value $(1-\beta)$ is known as the *power* of the test, indicating how *correctly* we can reject H_0 when it is actually false. Again, consider the case of hospital workers subject to a radiation exposure. Given the current estimate $S = s$ of standard deviation and the current sample size $n = n_1$, the t-statistic $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ can be approximated by $N(\delta, 1)$ with $\delta = \dfrac{\mu - \mu_0}{s/\sqrt{n_1}}$.



① $H_0: \mu = \mu_0$ vs. $H_A: \mu \neq \mu_0$

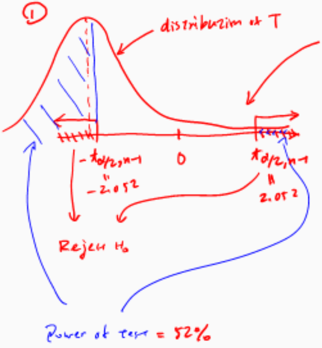
② $H_0: \mu = \mu_0$ vs. $H_A: \mu < \mu_0$

Example

Suppose that the true population mean is $\mu = 5.1$ (versus the value $\mu_0 = 5.4$ in our hypotheses). Then we can calculate the power of the test with $\delta \approx -2.11$ as follows.

How to calculate the power of test?

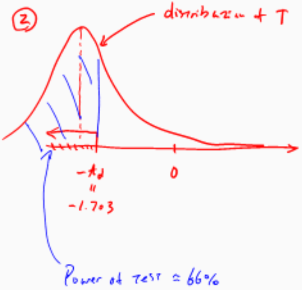
①



$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$= \underbrace{\underbrace{\frac{\bar{x} - \mu}{s/\sqrt{n}}}_{\text{T-distribution} \approx N(0,1)} + \underbrace{\frac{\mu - \mu_0}{s/\sqrt{n}}}_{\delta}}_{N(\delta,\,1)}$$

②



Example

Suppose that the true population mean is $\mu = 5.1$ (versus the value $\mu_0 = 5.4$ in our hypotheses). Then we can calculate the power of the test with $\delta \approx -2.11$ as follows.

$H_A: \mu \neq \mu_0$

1. In the two-sided hypothesis testing, we reject H_0 when $|T| > t_{0.025,27} = 2.052$. Therefore, the power of the test is $K(5.1) = P(|T| > 2.052 \mid \mu = 5.1) \approx 0.523$

   $N(\delta, 1)$

   $H_A: \mu < \mu_0$

2. In the one-sided hypothesis testing, we reject H_0 when $T < -t_{0.05,27} = -1.703$. Therefore, the power of the test is $K(5.1) = P(T < -1.703 \mid \mu = 5.1) \approx 0.658$

   $N(\delta, 1)$

This explains why we could not reject H_0 in the two-sided hypothesis testing. Our chance to detect the falsehood of H_0 is only 52%, while we have 66% of the chance in the one-sided hypothesis testing.

Effect of sample size.

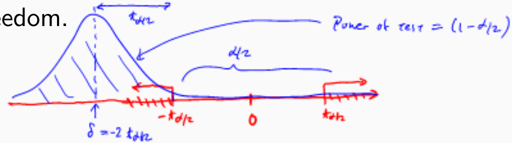
For a fixed significance level α of your choice, the power of the test increases as the sample size n increases. In the two-sided hypothesis testing discussed above, we could recommend to collect additional data to increase the power of the test. But how many additional data do we need? Here is one possible way to calculate a desirable sample size n: In the two-sided hypothesis testing, the power $(1-\beta)$ of the test is approximated by

$$T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \underbrace{\frac{\bar{x}-\mu}{s/\sqrt{n}}}_{Y} + \underbrace{\frac{\mu-\mu_0}{s/\sqrt{n}}}_{\delta} = Y + \delta$$

t-distribution

$$\begin{aligned} P(|T| > t_{\alpha/2,n-1}) &\approx P(Y < -t_{\alpha/2,n-1} - \delta) + P(Y > t_{\alpha/2,n-1} - \delta) \\ &\geq P(Y > t_{\alpha/2,n-1} - |\delta|) \end{aligned}$$

with a random variable Y having the t-distribution with $(n-1)$ degrees of freedom.



Effect of sample size, continued.

$$\frac{|\mu-\mu_0|}{s/\sqrt{n}} = 2t_{\alpha/2} \Rightarrow \sqrt{n} = \frac{(2t_{\alpha/2})s}{|\mu-\mu_0|}$$

Given the current estimate $S = s$ of standard deviation and the current sample size n_1, we can achieve the power $(1-\alpha/2)$ of the test by increasing a total sample size n and consequently satisfying $\boxed{|\delta| \geq 2t_{\alpha/2,n_1-1}.}$ In the above example of radiation exposure of hospital workers, such size n can be calculated as

$$n \geq \left(\frac{2t_{\alpha/2,n_1-1}s}{|\mu-\mu_0|}\right)^2$$

$$= \left(\frac{2t_{0.025,27} \times 0.7524}{|5.1-5.4|}\right)^2 = 105.9.$$

Comparison of two populations.

We often want to compare two populations on the basis of experiment. For example, a researcher wants to test the effect of his drug on blood pressure. In any treatment, an improvement could have been due to the *placebo effect* when the subject believes that he or she has been given an effective treatment. To protect against such biases, the study should consider (i) the use of a *control group* in which the subjects are given a placebo, and an *experimental group* in which the subjects are treated with the new drug, (ii) the *randomization* by assigning the subjects between the control and the exprimental groups randomly, and (iii) a *double-blind* experiment by concealing the nature of treatment from the subjects and the person taking measurements.

Hypothesis test.

This becomes the hypothesis testing problem

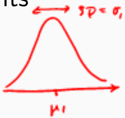
$$H_0: \ \mu_1 = \mu_2 \quad \text{versus} \quad H_A: \ \mu_1 \neq \mu_2.$$

where μ_1 and μ_2 are the respective population means of the control and the experimental groups

Normal assumption.

As a result of experiment, we typically obtain the measurements

$$X_1, \ldots, X_n \quad \sim N(\mu_1, \sigma_1^2)$$



of the subjects from the control group, and the measurements

$$Y_1, \ldots, Y_m \quad \sim N(\mu_2, \sigma_2^2)$$

of the subjects from the experimental group. Then it is usually assumed that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent and normally distributed with (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively. Even when they are not normally distributed, large sample sizes $(n, m \geq 30)$ ensure that the tests are appropriate via the central limit theorem.

Pooled variance procedure.

Let S_x and S_y be the sample standard deviations constructed from $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively. When it is reasonable to assume $\sigma^2 =$ "$\sigma_1^2 = \sigma_2^2$," we can construct the *pooled sample variance*

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

The test statistic

$$\boxed{T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}}$$

has the t-distribution with $(n+m-2)$ degrees of freedom under the null hypothesis H_0. Thus, we reject the null hypothesis H_0 with significant level α when the observed value t of T satisfies $|t| > t_{\alpha/2, n+m-2}$.

$\mu_1 = \mu_2$

Assumed

$X_1, \dots, X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_m \sim N(\mu_2, \sigma^2)$

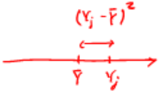
$\Downarrow$

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \qquad \bar{Y} = \frac{Y_1 + \dots + Y_m}{m}$$

$$S_x^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \qquad S_y^2 = \frac{\sum_{j=1}^{m} (Y_j - \bar{Y})^2}{m-1}$$

degree of freedom (df)

d.f.



$$S_p^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{j=1}^{m} (Y_j - \bar{Y})^2}{n+m-2} = \frac{(n-1) S_x^2 + (m-1) S_y^2}{n+m-2}$$

df

Pooled variance

$$\bar{X} \sim (\mu_1, \frac{\sigma^2}{n})$$

$$\bar{Y} \sim (\mu_2, \frac{\sigma^2}{m})$$

$$V_x = \frac{(n-1)}{\sigma^2} S_x^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2\text{-dist. with } (n-1) \text{ d.f.}$$

$$V_y = \frac{(m-1)}{\sigma^2} S_y^2 = \frac{1}{\sigma^2} \sum_{j=1}^{m} (Y_j - \bar{Y})^2 \sim \quad \text{"} \quad (m-1) \text{ d.f.}$$

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \sigma^2(\tfrac{1}{n} + \tfrac{1}{m}))$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(-\bar{Y})$$
$$= Var(\bar{X}) + (-1)^2 Var(\bar{Y})$$

$$\underbrace{V_x + V_y}_{\parallel} \sim \chi^2\text{-dist. with } (n+m-2) \text{ d.f}$$

$$\frac{(n+m-2)}{\sigma^2} S_p^2 = V$$

when $H_0: \mu_1 = \mu_2$

$$\bar{X} - \bar{Y} \sim N(0, \sigma^2(\tfrac{1}{n} + \tfrac{1}{m}))$$

$\Downarrow$

$$\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$$

Under H_A

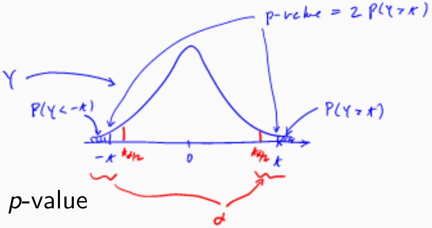
$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

follows t-dist.

$$T = \frac{N(0,1)}{\sqrt{\chi^2/df}} = \frac{\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{V/(n+m-2)}}$$

$$= \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Pooled variance procedure, continued.



Alternatively we can compute the p-value

$$p^* = 2 \times P(Y \geq |t|)$$

with Y having a t-distribution with $(n + m - 2)$ degrees of freedom, and reject H_0 when $p^* < \alpha$.

Confidence interval.

The following table shows the corresponding confidence interval of the population mean difference $\mu_1 - \mu_2$, when your null hypothesis H_0 is rejected.

Test	$(1-\alpha)$-level confidence interval for $\mu_1 - \mu_2$
$H_A : \mu_1 \neq \mu_2$.	$\left(\bar{X} - \bar{Y} - t_{\alpha/2,n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}},\ \bar{X} - \bar{Y} + t_{\alpha/2,n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right)$
$H_A : \mu_1 > \mu_2$.	$\left(\bar{X} - \bar{Y} - t_{\alpha,n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}},\ \infty\right)$
$H_A : \mu_1 < \mu_2$.	$\left(-\infty,\ \bar{X} - \bar{Y} + t_{\alpha,n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right)$

Example

Suppose that we consider the significant level $\alpha = 0.01$, and that we have obtained $\bar{X} = 80.02$ and $S_x = 0.024$ from the control group of size $n = 13$, and $\bar{Y} = 79.98$ and $S_y = 0.031$ from the experimental group of size $m = 8$. Here we have assumed that $\sigma_1^2 = \sigma_2^2$.

Example

Suppose that we consider the significant level $\alpha = 0.01$, and that we have obtained $\bar{X} = 80.02$ and $S_x = 0.024$ from the control group of size $n = 13$, and $\bar{Y} = 79.98$ and $S_y = 0.031$ from the experimental group of size $m = 8$. Here we have assumed that $\sigma_1^2 = \sigma_2^2$. ← What if it is not?

Then we can compute the square root $S_p = 0.027$ of the pooled sample variance S_p^2, and the test statistic

$$T = \frac{80.02 - 79.98}{0.027\sqrt{\frac{1}{13} + \frac{1}{8}}} \approx 3.33.$$

Thus, we can obtain $p^* = 2 \times P(Y \geq 3.33) \approx 0.0035 < 0.01$, and reject H_0. We conclude that the two population means are significantly different. And the 99% confidence interval for the mean difference is $(0.006, 0.074)$.

General procedure.

$\bar{X} - \bar{Y} \sim N(\underbrace{\mu_1 - \mu_2}_{\substack{\shortparallel \\ 0 \text{ under } H_0}}, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$

When "$\sigma_1^2 \neq \sigma_2^2$," under the null hypothesis H_0 the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \overset{\text{approx.}}{\sim} \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1) \text{ under } H_0$$

has approximately the t-distribution with ν degree of freedom, where ν is the nearest integer to

$$\frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{S_x^4}{n^2(n-1)} + \frac{S_y^4}{m^2(m-1)}}.$$

Thus, we reject the null hypothesis H_0 with significant level α when the observed value t of T satisfies $|t| > t_{\alpha/2,\nu}$.

General procedure, continued.

Alternatively we can compute the p-value

$$p^* = 2 \times P(Y \geq |t|)$$

with Y having a t-distribution with ν degrees of freedom, and reject H_0 when $p^* < \alpha$.

Confidence interval.

The following table shows the corresponding confidence interval of the population mean difference $\mu_1 - \mu_2$, when your null hypothesis H_0 is rejected.

Test	$(1-\alpha)$-level confidence interval for $\mu_1 - \mu_2$
$H_A : \mu_1 \neq \mu_2.$	$\left(\bar{X} - \bar{Y} - t_{\alpha/2,\nu}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}},\ \bar{X} - \bar{Y} + t_{\alpha/2,\nu}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\right)$
$H_A : \mu_1 > \mu_2.$	$\left(\bar{X} - \bar{Y} - t_{\alpha/2,\nu}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}},\ \infty\right)$
$H_A : \mu_1 < \mu_2.$	$\left(-\infty,\ \bar{X} - \bar{Y} + t_{\alpha/2,\nu}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\right)$

Example

Suppose that we consider the significant level $\alpha = 0.01$, and that we have obtained $\bar{X} = 80.02$ and $S_x = 0.024$ from the control group of size $n = 13$, and $\bar{Y} = 79.98$ and $S_y = 0.031$ from the experimental group of size $m = 8$ as before.

Example

Suppose that we consider the significant level $\alpha = 0.01$, and that we have obtained $\bar{X} = 80.02$ and $S_x = 0.024$ from the control group of size $n = 13$, and $\bar{Y} = 79.98$ and $S_y = 0.031$ from the experimental group of size $m = 8$ as before.

The test statistic $T \approx 3.12$, and

$$\frac{\left(\frac{(0.024)^2}{13} + \frac{(0.031)^2}{8}\right)^2}{\frac{(0.024)^4}{(13)^2(12)} + \frac{(0.031)^4}{(8)^2(7)}} \approx 12.15;$$

thus, we obtain $\nu = 12$ and $t_{0.005,12} \approx 3.055$. Since $T > t_{0.005,12}$, we still reject H_0. Alternatively we can obtain
$p^* = 2 \times P(Y \geq 3.12) \approx 0.0089 < 0.01$, and conclude that the difference is significant.

Inference on proportions.

Categorical data analysis:

outcomes	c_1	c_2	$\cdots$	c_k
probability	p_1	p_2		p_k

$$\longrightarrow \underbrace{x_1, x_2, \ldots, x_n}_{\substack{\text{categorical values} \\ c_1, c_2, \ldots, c_k}}$$

In experiments on pea breeding, Mendel observed the different kinds of seeds obtained by crosses from plants with round yellow seeds and plants with wrinkled green seeds. Possible types of progeny were: "round yellow", "wrinkled yellow", "round green", and "wrinkled green." When the data values recorded $x_1, \ldots, x_n$ takes several types, or categories, we call them the *categorical data*.

Bernoulli trials:

outcomes	success	failure
probability	p	$1-p$

$$\longrightarrow \underbrace{x_1, x_2, \ldots, x_n}_{\text{success or failure}}$$

Point estimate.

Bernoulli trials: $X_i = \begin{cases} 1 & \text{with success} \\ 0 & \text{with failure} \end{cases}$ $\Rightarrow$ $X = X_1 + \cdots + X_n = \#$ of successes $\sim$ Binomial with (n, p)

$E[X] = np$ $\quad Var(X) = np(1-p)$

Let X be the number of observations for a particular type in categorical data of size n, and let p be the *population proportion* of this type (that is, the probability of occurrence of this type). Then the random variable X has the binomial distribution with parameter (n, p). And the point estimate of the population proportion p is

$$\hat{p} = \frac{X}{n}.$$

We can easily see that

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = p$$

Thus, $\hat{p}$ is an unbiased estimate of p.

Point estimate, continued.

$X = \sum_{i=1}^{n} X_i \overset{\text{approx}}{\sim} N(np, np(1-p))$

$\hat{p} = \frac{X}{n} \overset{\text{approx}}{\sim} N(p, \frac{p(1-p)}{n})$

$Var(\hat{p}) = Var((\frac{1}{n})X) = \frac{p(1-p)}{n}$

Recall by the central limit theorem that we have approximately

$$X \overset{\text{approx}}{\sim} N(np, np(1-p))$$

when n is large. Then the point estimate $\hat{p}$ is approximately distributed as the normal distribution with parameter $(p, \frac{p(1-p)}{n})$.

Hypothesis test.

Claim: $p < p_0 = 0.05$ $\xleftarrow{\text{opposite}}\!\!\rightarrow$ $H_0 : p \geq p_0$

$\downarrow$ Has to be equal "=" for H_0

$H_0 \quad p = p_0$

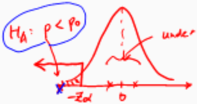
Suppose that the vaccine can be approved for widespread use if it can be established that the probability p of serious adverse reaction is less than p_0. Then the hypothesis testing problem becomes

$$H_0 : \; p = p_0 \quad \text{versus} \quad H_A : \; p < p_0. \tag{8.2}$$

Let X be the number of participants who suffer an adverse reaction among n participants. Then, the random variable X has the binomial distribution with parameter (n, p) and is approximated by the normal distribution with parameter $(np, np(1-p))$ when n is large [that is, to satisfy $np > 5$ and $n(1-p) > 5$].

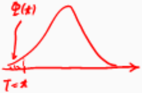
Assume $H_0: p = p_0 \Rightarrow X \overset{\text{approx}}{\sim} N(\underbrace{np_0}_{\mu}, \underbrace{np_0(1-p_0)}_{\sigma^2})$

The critical point of the standard normal distribution, denoted by z_α, is defined as the value satisfying $P(Z > z_\alpha) = \alpha$ where Z is a standard normal random variable. Since the normal distribution is symmetric, it implies that $P(Z < -z_\alpha) = \alpha$. When $np_0 > 5$ and $n(1 - p_0) > 5$,



$$\frac{X - \mu}{\sigma} = T = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}} \overset{\text{approx.}}{\sim} N(0, 1) \tag{8.3}$$

is used for the test statistic. Then we can reject H_0 in (8.2) with significance level α if the value t of the test statistic T satisfies $t < -z_\alpha$. Equivalently, we can proceed to construct the p-value $p^* = P(Z < t) = \Phi(t)$, and reject H_0 when $p^* < \alpha$. Since the consideration of continuity correction improves the accuracy,



$$T = \frac{X - np_0 + 0.5}{\sqrt{np_0(1 - p_0)}}$$

may be used instead.

Confidence interval.

in favor of $H_A: p < p_0$

When H_0 is rejected, we want to further investigate the confidence interval for the population proportion p which corresponds to the result of hypothesis test. We have the point estimate $\hat{p} = X/n$. Then the two different formulas

$$\left(0,\ \frac{X + z_\alpha^2/2 + z_\alpha\sqrt{X(n-X)/n + z_\alpha^2/4}}{n + z_\alpha^2}\right) \tag{8.4}$$

$$\left(0,\ \hat{p} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \tag{8.5}$$

can be used for the confidence interval of level α. Although Formula (8.4) is known to be more accurate, Formula (8.5) may be used in most of our problems since it is easier to calculate.

Example

Suppose that $p_0 = 0.05$ is required, and that the significance level $\alpha = 0.05$ is chosen. And the study shows that $X = 4$ adverse reactions are found out of $n = 155$ participants.

$H_0: p = 0.05$ vs. $H_A: p < 0.05$

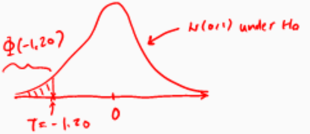
Example

Suppose that $p_0 = 0.05$ is required, and that the significance level $\alpha = 0.05$ is chosen. And the study shows that $X = 4$ adverse reactions are found out of $n = 155$ participants.

Note that $(0.05)(155) = 7.75 > 5$ and $(0.95)(155) = 147.25 > 5$. Thus, we have

$$T = \frac{4 - (155)(0.05) + 0.5}{\sqrt{(155)(0.05)(0.95)}} \approx -1.20 \quad \text{and} \quad p^* = \Phi(-1.20) \approx 0.115$$

We can also obtain the point estimate $\hat{p} \approx 0.0258$ and the 95% confidence interval $(0,\ 0.0562)$ by using (8.4) [we get $(0,\ 0.0467)$ if we use (8.5)]. Since $p^* \geq 0.05$, we cannot reject the null hypothesis. Thus, it is not advisable that the vaccine be approved as the result of this study.



Sample size calculations.

We always guarantee the possibility of incorrectly rejecting H_0 when H_0 is true—Type I error, say, to be less than 5% of the chance. But, at the same time we sacrifice the power of detecting the falsehood of H_0 when H_0 is false—power of the test. In order for the hypothesis testing problem

$$H_0:\ p = p_0 \quad \text{versus} \quad H_A:\ p < p_0,$$

to achieve the power $(1-\beta)$ of the test, we need a sample of size

$$n \geq \left(\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)}}{p - p_0} \right)^2 . \tag{8.6}$$

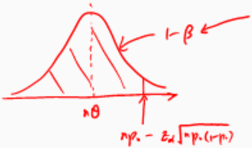
with $\theta = p$

T

$P\left(\frac{X - np_0}{\sqrt{np_0(1-p_0)}} < -z_\alpha\right) \geq 1-\beta$

$= P\left(X < np_0 - z_\alpha\sqrt{np_0(1-p_0)}\right)$

$X \sim N(n\theta, n\theta(1-\theta))$ if $p = \theta$



It happens if

$np_0 - z_\alpha\sqrt{np_0(1-p_0)} \geq \underbrace{n\theta + z_\beta\sqrt{n\theta(1-\theta)}}$

Sample size calculations: Example.

$H_0: p = 0.05$ vs. $H_A: p < 0.05$

$$P(\underset{\Uparrow}{T} < -z_\alpha \mid p = \theta) = P(\tilde{X} < \underbrace{\overset{155}{n}\overset{0.05}{p_0} - \overset{1.645 \text{ if } \alpha = 0.05}{z_\alpha}\sqrt{np_0(1-p_0)}}_{3.287} \mid p = \theta)$$

$$\frac{\tilde{X} - np_0}{\sqrt{np_0(1-p_0)}}$$

$$\tilde{X} \sim N(n\theta, n\theta(1-\theta)) = N(3.875, 3.778) \text{ if } \theta = 0.025$$

In the example of vaccine experiment, if the true population mean p is 0.025, then the power of the test is calculated as

$$K(0.025) = P(\text{“Reject } H_0\text{”} \mid p = 0.025) \approx P(\tilde{X} < 3.287) \approx 0.38$$

To increase the power of the test at least 0.8 $= (1-\beta)$ $\Downarrow$ $\beta = 0.2$, we need the sample size at least $n = 384$.

Summary.

Possible null hypotheses for the inference on population proportion are "$H_A: p \neq p_0$", "$H_A: p > p_0$", and "$H_A: p < p_0$". In either case we can use the test statistic Z in (8.3) if we do not make a "continuity correction." Then the corresponding testing procedures are summarized in the following table.

Test	When to reject	$(1-\alpha)$-level confidence interval
$H_A: p \neq p_0$	$\lvert Z \rvert > z_{\alpha/2}$	$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$
$H_A: p > p_0$	$Z > z_\alpha$	$\left(\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ 1\right)$
$H_A: p < p_0$	$Z < -z_\alpha$	$\left(0,\ \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

Summary, continued.

For the sample size calculation, use Formula (8.6) if the null hypothesis is either "$H_A:\ p > p_0$" or "$H_A:\ p < p_0$". When the null hypothesis is "$H_A:\ p \neq p_0$", the sample size n can be computed as

$$n \geq \left(\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta \sqrt{p(1-p)}}{p - p_0} \right)^2 .$$

Comparison of two proportions.

A researcher is interested in whether there is discrimination against women in a university. In terms of statistics this is the hypothesis testing problem

$$H_0: \ p_A = p_B \quad \text{versus } H_A: \ p_A > p_B$$

where p_A and p_B are the respective population proportions of men and women who are admitted to the university. The researcher decided to collect the data for graduate program in the university.

Point estimate.

Let X and Y be the respective numbers of men and women who are admitted to the graduate school.

$H_0 : p_A = p_B = p$

	Men	Women
Admit	X	Y
Deny	$n - X$	$m - Y$
Total	n	m

$X \sim N(np, np(1-p))$

$Y \sim N(mp, mp(1-p))$

$\Downarrow$

$\hat{p}_A = \frac{X}{n} \sim N(p, \frac{p(1-p)}{n})$

$\hat{p}_B = \frac{Y}{m} \sim N(p, \frac{p(1-p)}{m})$

$\Downarrow$

$\hat{p}_A - \hat{p}_B \sim N(0, p(1-p)(\frac{1}{n} + \frac{1}{m}))$

standardize

$Var(\hat{p}_A - \hat{p}_B) = Var(\hat{p}_A) + Var(\hat{p}_B)$

The test statistic is given by

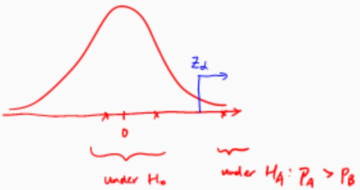
$$N(0,1) \text{ under } H_0! \leftarrow Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

where $\hat{p}_A = X/n$ and $\hat{p}_B = Y/m$ are the point estimates of p_A and p_B, and

$$\hat{p} = \frac{X+Y}{n+m}$$

is called a *pooled* estimate of the common population proportion.

Hypothesis test.



Under the null hypothesis, the probability that $Z > z_\alpha$ becomes approximately less than α. Thus, we reject H_0 when the observed value z of Z satisfies $z > z_\alpha$. Or, equivalently we can reject $p^* = 1 - \Phi(z) < \alpha$.

Confidence interval.

We may want to further investigate the confidence interval for the difference $p_A - p_B$. Having constructed the hypothesis test problems "$H_A: p_A \neq p_B$", "$H_A: p_A > p_B$", or "$H_A: p_A < p_B$", the following table shows the corresponding testing procedure and the confidence interval.

Test	Rejection	$(1-\alpha)$-level confidence interval
$H_A: p_A \neq p_B$	$\lvert z \rvert > z_{\alpha/2}$	$\left(\hat{p}_A - \hat{p}_B - z_{\alpha/2}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}},\ \hat{p}_A - \hat{p}_B + z_{\alpha/2}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}},\right)$
$H_A: p_A > p_B$	$z > z_\alpha$	$\left(\hat{p}_A - \hat{p}_B - z_\alpha\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}},\ 1\right)$
$H_A: p_A < p_B$	$z < -z_\alpha$	$\left(-1,\ \hat{p}_A - \hat{p}_B + z_\alpha\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n} + \frac{\hat{p}_B(1-\hat{p}_B)}{m}}\right)$

Example

The following table classifies the applications for the graduate school according to admission status and sex.

	Men	Women	Total
Admit	97	40	137
Deny	263	42	305
Total	360	82	442

Example

The following table classifies the applications for the graduate school according to admission status and sex.

	Men	Women	Total
Admit	97	40	137
Deny	263	42	305
Total	360	82	442

No evidence for

$H_A: p_A > p_B$

We have $\hat{p}_A = 97/360 \approx 0.269$, $\hat{p}_B = 40/82 \approx 0.488$, and $\hat{p} = 137/442 \approx 0.310$. And we can obtain

$$Z = \frac{0.269 - 0.488}{\sqrt{(0.31)(0.69)(1/360 + 1/82)}} \approx -3.87 \quad \text{and} \quad p^* = 1 - \Phi(-3.87) \approx 0.99$$

Thus, we cannot reject H_0, indicating that there is no discrimination against women in this particular graduate program. In fact, the alternative hypothesis "$H_A : p_A < p_B$" will be established in this example.

Goodness of fit.

In the experiment on pea breeding, Mendel observed the different kinds of seeds obtained by crosses from plants with round yellow seeds and plants with wrinkled green seeds. Possible types of progeny were: "round yellow (RY)", "wrinkled yellow (WY)", "round green (RG)", and "wrinkled green (WG)." And Mendel's theory predicted the associated probabilities of occurrence as follows.

	RY	WY	RG	WG
Probabilities	9/16	3/16	3/16	1/16
	p_1	p_2	p_3	p_4

We want to test whether the data from n observation is consistent with his theory—goodness of fit test, in which the statement of null hypothesis becomes "the model is valid."

Chi-square test.

In general, each observation is classified into one of k categories or "cells," which results in the *cell frequencies*

$$c_1, \ldots, c_k \qquad X_1, \ldots, X_k. \longrightarrow x_1 + x_2 + \cdots + x_k = n$$

The goodness of fit to a particular model can be assessed by comparing the observed cell frequencies $X_1, \ldots, X_k$ with the expected cell frequencies

$$E_1, \ldots, E_k, \qquad np_1, \ldots, np_k$$

which are predicted from the model. The discrepancy between the data and the model can be measured by the *Pearson's chi-square statistic*

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}. \overset{\text{approx}}{\sim} \chi^2\text{-dist.}$$

Why $\sum_{i=1}^{k} \frac{(x_i - E_i)^2}{E_i}$ follows χ^2-distribution?

Consider $k=2$:

	C_1	C_2
Probability	p_1	$p_2 = (1-p_1)$

$$X_1 \sim \text{Binomial}(n, p_1) \overset{\text{approx}}{\sim} N(np_1, np_1(1-p_1))$$

$$\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \overset{\text{approx}}{\sim} N(0,1)$$

$$\left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2 \overset{\text{approx}}{\sim} \chi^2\text{-dist}$$

$$\|$$

$$\frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \left(\frac{1}{np_1} + \frac{1}{n(1-p_1)}\right)(X_1 - np_1)^2$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X_1 - E_1)^2}{E_1} + \frac{(X_2 - E_2)^2}{E_2}$$

$$= \sum_{i=1}^{2} \frac{(X_i - E_i)^2}{E_i}$$

$$(X_1 - np_1)^2 = [(n - X_2) - n(1-p_2)]^2$$

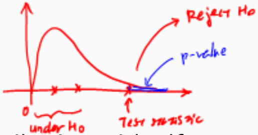
$$= [-X_2 + np_2]^2 = (X_2 - np_2)^2$$

Chi-square test, continued.

Under the null hypothesis (that is, assuming that the model is correct), the distribution of Pearson's chi-square χ^2 is approximated by the chi-square distribution with

$$df = (\text{number of cells}) - 1 - (\text{number of parameters in the model})$$

degrees of freedom. Therefore, if you observe that $\chi^2 = x$ and $x > \chi^2_{\alpha,df}$, then we can reject the null hypothesis, casting doubt on the validity of the model. Or, by computing the p-value



$$p^* = P(X > x)$$

with a random variable X having the chi-square distribution with df degrees of freedom, equivalently we can reject the null hypothesis when $p^* < \alpha$.

Example

In the experiment of pea breeding, we have obtained the data as in the following table.

	RY	WY	RG	WG
Frequencies	315	101	108	32

Example

In the experiment of pea breeding, we have obtained the data as in the following table.

	RY	WY	RG	WG
Frequencies	315	101	108	32

With the total number of observations $n = 556$, the expected cell frequencies from the Mendel's theory can be calculated as

	RY	WY	RG	WG
Expected frequencies	312.75	104.25	104.25	34.75

We can compute the Pearson's chi-square $\chi^2 = 0.47$. Since the Mendel's model has no parameter, the chi-square distribution has $3 = (4 - 1)$ degrees of freedom and we get the p-value $p^* = 0.925$. Thus, there is little reason to doubt the Mendel's theory on the basis of Pearson's chi-square test.

Test of independence.

Consider again the study of discrimination against women in university admission. In the study, there are two characteristics: men or women; admitted or denied. The researcher wanted to know whether such characteristics are linked or independent. For such a study, we take a random sample of size n from the population, which is summarized in the *contingency table*

	Men	Women	Total
Admit	X_{11}	X_{12}	$X_{1\cdot}$
Deny	X_{21}	X_{22}	$X_{2\cdot}$
Total	$X_{\cdot 1}$	$X_{\cdot 2}$	$n = X_{\cdot\cdot}$

Test of independence, continued.

The statement of null hypothesis becomes "the two characteristics are independent." Under the null hypothesis, the expected frequencies for the contingency table can be given by

Expected number E_{11}

	Men	Women	Total
Admit	np_1q_1	np_1q_2	np_1
Deny	np_2q_1	np_2q_2	$np_2 = n(1-p_1)$
Total	nq_1	$nq_2 = n(1-q_1)$	n

The point estimates of p_1, p_2, q_1, and q_2 are $\hat{p}_1 = X_{1\cdot}/n$, $\hat{p}_2 = X_{2\cdot}/n$, $\hat{q}_1 = X_{\cdot 1}/n$, and $\hat{q}_2 = X_{\cdot 2}/n$. With these point estimates, the chi-square statistic is

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(X_{ij} - X_{i\cdot}X_{\cdot j}/n)^2}{X_{i\cdot}X_{\cdot j}/n}, = \sum_{i,j=1}^{2} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2\text{-dist.}$$

and the degree of freedom is $(4 - 1 - 2) = 1$.

Test of independence: Example.

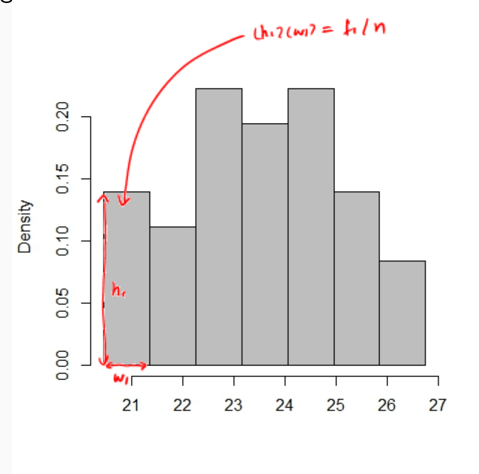
By using the same data as before, we can obtain the chi-square statistic

$$\chi^2 = \frac{[97 - (137)(360)/(442)]^2}{(137)(360)/(442)} + \frac{[40 - (137)(82)/(442)]^2}{(137)(82)/(442)} + \frac{[263 - (305)(360)/(442)]^2}{(305)(360)/(442)} + \frac{[42 - (305)(82)/(442)]^2}{(305)(82)/(442)} \approx 14.89,$$

and the p-value $p^* = 0.0001$. Thus, the null hypothesis is rejected at any reasonable level, indicating that the two characteristics are somewhat dependent.

Histogram.

The data points $x_1, \ldots, x_n$ are typically considered as the observed values of random variables $X_1, \ldots, X_n$ having a common probability distribution $f(x)$. To judge the quality of data, it is useful to envisage a *population* by the following graphical representation, called *histogram*.



Relative frequency.

Consider the intervals

$$[20.45, 21.35), [21.35, 22.25), [22.25, 23.15), \ldots, [25.85, 26.75)$$

$w = 0.9$

Then the number of observations f_i in the i-th interval becomes a *sample frequency*, and the density

$$h_i = \frac{f_i}{n \times (\text{width of the } i\text{-th interval})}.$$

forms the height of the i-th rectangle above the i-th interval in the histogram. When the width of each interval is equally chosen, the width w is called *bandwidth* and the height h_i becomes

$$h_i = \frac{f_i}{n \times w}. = \frac{f_i / n}{w}$$

Stem and leaf plot.

A *stem and leaf plot* is much like a histogram except it portrays a data set itself. The leading digit(s) of the data values become stems, which split the trailing digit(s) as leaves. The trailing digits are rounded down to a single digit if necessary.

5 data points [20.45, 21.35)

20.5	20.7	20.8	21.0	21.0	21.4
21.5	22.0	22.1	22.5	22.6	22.6
22.7	22.7	22.9	22.9	23.1	23.3
23.4	23.5	23.6	23.6	23.6	23.9
24.1	24.3	24.5	24.5	24.8	24.8
24.9	24.9	25.1	25.1	25.2	25.6
25.8	25.9	26.1	26.7		

$\Longrightarrow$

20	578
21	0045
22	015667799
23	13456669

Median and quartiles.

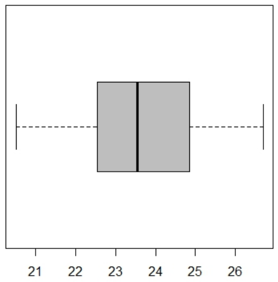
The *median* is the value of the "middle" data point, and defined by

$$\begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd;} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } n \text{ is even.} \end{cases}$$

For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is $(4+7)/2 = 5.5$. The 25-percentile is the value indicating that 25% of the observations takes values smaller than the value. Similarly, we can define 50-percentile, 75-percentile, and so on. Note that 50-percentile is the median. We call 25-percentile the *lower sample quartile* and 75-percentile the *upper sample quartile*.

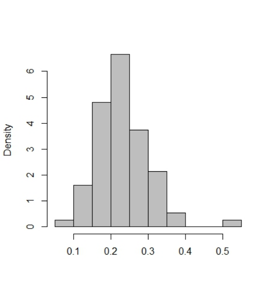
Boxplot.

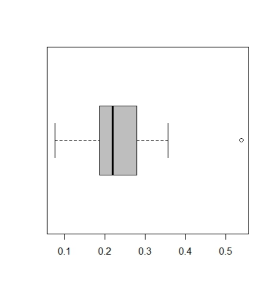
A box is drawn stretching from the lower sample quartile (the 25-percentile) to the upper quartile (the 75-percentile). The median is shown as a line across the box. Therefore $1/4$ of the distribution is between this line and the right of the box and $1/4$ of the distribution is between this line and the left of the box. Vertical lines (dotted), called "whiskers," stretch out from the ends of the box to the largest and smallest data.



Outliers.

Graphical presentations can be used to identify "odd-looking" value which does not fit in with the rest of the data. Such a value is called an *outlier*. In many cases an outlier is discovered to be a misrecorded data value, or represents some special condition that was not in effect when the data were collected. In the histogram and boxplot below, the value in far right appears to be quite separate from the rest of the data, and can be considered to be an outlier.





Exercises

Problem

An experimenter is interested in the hypothesis testing problem

$$H_0:\ \mu = 3.0mm \quad \text{versus } H_A:\ \mu \neq 3.0mm,$$

where μ is the population mean of thickness of glass sheets. Suppose that a sample of $n = 21$ glass sheets is obtained and their thicknesses are measured.

1. *For what values of the t-statistic does the experimenter accept the null hypothesis with a size $\alpha = 0.10$?*
2. *For what values of the t-statistic does the experimenter reject the null hypothesis with a size $\alpha = 0.01$?*

Suppose that the sample mean $\bar{X} = 3.04mm$ and the sample standard deviation is $S = 0.124mm$. Is the null hypothesis accepted or rejected with $\alpha = 0.10$? With $\alpha = 0.01$?

Problem

A machine is set to cut metal plates to a length of 44.350mm. The length of a random sample of 24 metal plates have a sample mean of $\bar{X} = 44.364$*mm and a sample standard deviation of* $S = 0.019$*mm. Is there any evidence that the machine is miscalibrated?*

Problem

An experimenter is interested in the hypothesis testing problem

$$H_0:\ \mu = 0.065 \quad \textit{versus} \quad H_A:\ \mu > 0.065$$

where μ *is the population mean of the density of a chemical solution. Suppose that a sample of* $n = 31$ *bottles of the chemical solution is obtained and their densities are measured.*

1. *For what values of the t-statistics does the experimenter accept the null hypothesis with a size* $\alpha = 0.10$*?*
2. *For what values of the t-statistics does the experimenter reject the null hypothesis with a size* $\alpha = 0.01$*?*

Suppose that the sample mean $\bar{X} = 0.0768$ *and the sample standard deviation is* $S = 0.0231$*. Is the null hypothesis accepted or rejected with* $\alpha = 0.10$*? With* $\alpha = 0.01$*?*

Problem

A chocolate bar manufacturer claims that at the time of purchase by a consumer the average age of its product is no more than 120 days. In an experiment to test this claim a random sample of 26 chocolate bars are found to have ages at the time of purchase with a sample mean of $\bar{X} = 122.5$ days and a sample standard deviation of $S = 13.4$ days. With this information how do you feel about the manufacturer's claim?

Problem

Measurement of suspended particles in $\mu g/m^3$ can be made for air quality monitoring. Let μ_1 and μ_2 be the average concentration of suspended particles in the city center of Melbourne and Houston, respectively. Using $n_1 = 13$ observations for Melbourne and $n_2 = 16$ observations for Houston, we test

$$H_0 : \mu_1 = \mu_2 \quad \textit{versus} \quad H_A : \mu_1 < \mu_2$$

1. *Assuming the equal variances, find the critical region with $\alpha = 0.05$.*
2. *Suppose that $\bar{X} = 72.9$, $S_1 = 25.6$, $\bar{Y} = 81.7$ and $S_2 = 28.3$ for Melbourne and Houston, respectively. Then calculate the test statistic, and state the conclusion.*

Problem

14% of drivers used a seat belt, and an advertising campaign was conducted to increase this proportion. Two months after the campaign, 104 drivers ouf of a random sample of 590 drivers were wearing a seat belt.

1. *Define the null hypothesis and alternative hypothesis using the proportion p of driviers using a seat belt after the campaign.*
2. *Find the critical region with* $\alpha = 0.01$.
3. *Calculate the p-value, and state the conclusion.*

Answers to exercises

Problem 10.

(a) For the t-statistic T, we fail to reject the null hypothesis if $|T| \leq t_{0.05,20} = 1.725$.

(b) For the t-statistic T, we reject the null hypothesis if $|T| > t_{0.005,20} = 2.845$.

We obtain $T = \frac{3.04-3.0}{0.124/\sqrt{21}} \approx 1.478$, and therefore, we fail to reject the null hypothesis with $\alpha = 0.10$. Since the p-value must be greater than 0.10, we clearly fail to reject the null hypothesis with $\alpha = 0.01$.

Problem 11.

We can set the hypotheses

$$H_0 : \mu = 44.350 \quad \text{versus} \quad H_A : \mu \neq 44.350$$

Since $T = \frac{44.364 - 44.350}{0.019/\sqrt{24}} \approx 3.61 > t_{0.005,23} = 2.807$, we reject the null hypothesis with $\alpha = 0.01$, and find evidence that the machine is miscalibrated.

Problem 12.

(a) For the t-statistic T, we fail to reject the null hypothesis if $T \leq t_{0.1,30} = 1.310$.

(b) For the t-statistic T, we reject the null hypothesis if $T > t_{0.01,30} = 2.457$.

We obtain $T = \frac{0.0768-0.065}{0.0231/\sqrt{31}} \approx 2.844$, and therefore, we reject the null hypothesis with $\alpha = 0.01$. Since the p-value must be less than 0.01, we are clearly able to reject the null hypothesis with $\alpha = 0.10$.

Problem 13.

We can set the hypotheses

$$H_0 : \mu = 120 \quad \text{versus} \quad H_A : \mu > 120$$

Since $T = \frac{122.5-120}{13.4/\sqrt{26}} \approx 0.951 < t_{0.1,25} = 1.316$, we fail to reject the null hypothesis with $\alpha = 0.1$, and could not find sufficient evidence against the manufacturer's claim.

Problem 14.

(a) For the test statistic T, we can form the critical region $T < -t_{0.05,27} = -1.703$.

(b) We can calculate

$$S_p^2 = \frac{(12)(25.6)^2 + (15)(28.3)^2}{27} \approx 736.21$$
$$T = \frac{72.9 - 81.7}{\sqrt{736.21}\sqrt{1/13 + 1/16}} \approx -0.869$$

Thus, we fail to reject H_0, and therefore, we cannot find sufficient evidence that Melbourne has the lower concentration of suspended particles than Houston.

Problem 15.

(a) We define the hypotheses

$$H_0 : p = 0.14 \quad \text{versus} \quad H_A : p > 0.14$$

(b) For the test statistic Z we can form the critical region $Z > z_{0.01} = 2.326$.

(c) We obtain $Z = \frac{104-(590)(0.14)}{\sqrt{(590)(0.14)(0.86)}} \approx 2.54$, and therefore, calculate p-value by $1 - \Phi(2.54) = 0.0055$. Thus, we reject H_0, and therefore, we find evidence that the campaign was successful.