Lecture Stat 461-561
Review of Hypothesis Testing and p-values

AD

January 2007

Example of Hypothesis Testing: Suppose we want to know if exposure to asbestos is associated with lung disease. We could take some rats and randomly divide them in 2 groups: one group exposed to asbestos and the other unexposed. Then we compare the disease rate.

- **Null Hypothesis**: The disease rate is the same in the two groups.

Example of Hypothesis Testing: Suppose we want to know if exposure to asbestos is associated with lung disease. We could take some rats and randomly divide them in 2 groups: one group exposed to asbestos and the other unexposed. Then we compare the disease rate.

- **Null Hypothesis**: The disease rate is the same in the two groups.
- **Alternative Hypothesis**: The disease rate is not the same in the two groups.
  If the exposed group has a much higher rate of disease than the unexposed group, then we will reject the null hypothesis and conclude that the evidence favors the alternative hypothesis.

Example of Hypothesis Testing: Suppose we want to know if exposure to asbestos is associated with lung disease. We could take some rats and randomly divide them in 2 groups: one group exposed to asbestos and the other unexposed. Then we compare the disease rate.

- **Null Hypothesis**: The disease rate is the same in the two groups.
- **Alternative Hypothesis**: The disease rate is not the same in the two groups.
  If the exposed group has a much higher rate of disease than the unexposed group, then we will reject the null hypothesis and conclude that the evidence favors the alternative hypothesis.
- Formally we can partition the parameter space into two disjoint sets Θ_0 and Θ_1 and we test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

- Let X be a random variable and let $\mathcal{X}$ be the range of X.

- Let X be a random variable and let $\mathcal{X}$ be the range of X.
- We test an hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**.

- Let X be a random variable and let $\mathcal{X}$ be the range of X.
- We test an hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**.
- If $X \in R$, we reject H_0, otherwise we do not.

- Let X be a random variable and let $\mathcal{X}$ be the range of X.
- We test an hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection region**.
- If $X \in R$, we reject H_0, otherwise we do not.
- Usually the rejection region is of the form

$$R = \{x : T(x) > c\}$$

where T is a **test statistic** and c is a **critical value**.

- There are two types of errors

	Retain H_0	Reject H_0
H_0 true	$\times$	type I error
H_1 true	type II error	$\times$

- There are two types of errors

	Retain H_0	Reject H_0
H_0 true	$\times$	type I error
H_1 true	type II error	$\times$

- The **power function** of a test with rejection R is the **probability of rejecting** the hypothesis defined by

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

- There are two types of errors

	Retain H_0	Reject H_0
H_0 true	$\times$	type I error
H_1 true	type II error	$\times$

- The **power function** of a test with rejection R is the **probability of rejecting** the hypothesis defined by

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

- An ideal test would satisfy $\beta(\theta) = 0$ for $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for $\theta \in \Theta_1$.

- There are two types of errors

	Retain H_0	Reject H_0
H_0 true	$\times$	type I error
H_1 true	type II error	$\times$

- The **power function** of a test with rejection R is the **probability of rejecting** the hypothesis defined by

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

- An ideal test would satisfy $\beta(\theta) = 0$ for $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for $\theta \in \Theta_1$.
- The **size of a test** is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

It is the **largest possible probability of making an error of type I,** i.e. the maximum power under the null hypothesis.

- There are two types of errors

	Retain H_0	Reject H_0
H_0 true	$\times$	type I error
H_1 true	type II error	$\times$

- The **power function** of a test with rejection R is the **probability of rejecting** the hypothesis defined by

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

- An ideal test would satisfy $\beta(\theta) = 0$ for $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for $\theta \in \Theta_1$.
- The **size of a test** is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

It is the **largest possible probability of making an error of type I,** i.e. the maximum power under the null hypothesis.

- A **test has significance level** α if its size is less than or equal to α.

- Hypothesis of the form $\theta = \theta_0$ is a **simple hypothesis**.

- Hypothesis of the form $\theta = \theta_0$ is a **simple hypothesis**.
- Hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is a **composite hypothesis**.

- Hypothesis of the form $\theta = \theta_0$ is a **simple hypothesis**.
- Hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is a **composite hypothesis**.
- A **two-sided test** is of the form

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

- Hypothesis of the form $\theta = \theta_0$ is a **simple hypothesis**.
- Hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is a **composite hypothesis**.
- A **two-sided test** is of the form

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

- A one-sided test is of the the form

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0.$$

- We have not yet discussed how to set c.

- We have not yet discussed how to set c.
- Assuming $R = \{x : T(x) > c\}$ then if we choose the critical value c to satisfy

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta (X \in R)$$

then by construction we get a test of size α.

- **Example**. Let $X_1, ..., X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ where σ is known. We want to test $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$. Hence $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, +\infty)$.

- **Example**. Let $X_1, ..., X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ where σ is known. We want to test $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$. Hence $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, +\infty)$.
- Consider the test

$$\text{reject } H_0 \text{ if } \overline{X} > c.$$

- **Example**. Let $X_1, ..., X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ where σ is known. We want to test $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$. Hence $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, +\infty)$.
- Consider the test

$$\text{reject } H_0 \text{ if } \overline{X} > c.$$

- The rejection region is

$$R = \left\{ (x_1, ..., x_n) : T(x_1, ..., x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i > c \right\}.$$

- **Example**. Let $X_1, ..., X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ where σ is known. We want to test $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$. Hence $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, +\infty)$.
- Consider the test

$$\text{reject } H_0 \text{ if } \overline{X} > c.$$

- The rejection region is

$$R = \left\{ (x_1, ..., x_n) : T(x_1, ..., x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i > c \right\}.$$

- The power function is

$$\begin{aligned} \beta(\mu) &= \mathbb{P}_\mu\left(\overline{X} \in R\right) = \mathbb{P}_\mu\left(\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} > \frac{\sqrt{n}\left(c - \mu\right)}{\sigma}\right) \\ &= \mathbb{P}\left(Z > \frac{\sqrt{n}\left(c - \mu\right)}{\sigma}\right) \text{ where } Z \sim \mathcal{N}(0, 1) \\ &= 1 - \Phi\left(\frac{\sqrt{n}\left(c - \mu\right)}{\sigma}\right). \end{aligned}$$

- This function is increasing in μ, hence

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

- This function is increasing in μ, hence

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

- For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

- This function is increasing in μ, hence

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

- For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}.$$

- We reject when $\overline{X} > \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}$, equivalently when

$$\frac{\sqrt{n}\left(\overline{X} - 0\right)}{\sigma} > z_\alpha$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$ (i.e. $P(Z \geq z_\alpha) = \alpha$).

- Assume you have $X_i \overset{\text{i.i.d.}}{\sim} f\left(x\middle|\theta\right)$ then the Likelihood Ratio (LR) test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is defined by

$$\lambda\left(\mathbf{x}\right) = \frac{\sup_{\theta\in\Theta_0} L\left(\theta\middle|\mathbf{x}\right)}{\sup_{\theta\in\Theta} L\left(\theta\middle|\mathbf{x}\right)}$$

where $L\left(\theta\middle|\mathbf{x}\right)$ is the likelihood function

$$L\left(\theta\middle|\mathbf{x}\right) = \prod_{i=1}^{n} f\left(x_i\middle|\theta\right).$$

- Assume you have $X_i \overset{\text{i.i.d.}}{\sim} f\left(x\middle|\theta\right)$ then the Likelihood Ratio (LR) test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is defined by

$$\lambda\left(\mathbf{x}\right) = \frac{\sup_{\theta\in\Theta_0} L\left(\theta\middle|\mathbf{x}\right)}{\sup_{\theta\in\Theta} L\left(\theta\middle|\mathbf{x}\right)}$$

  where $L\left(\theta\middle|\mathbf{x}\right)$ is the likelihood function

$$L\left(\theta\middle|\mathbf{x}\right) = \prod_{i=1}^{n} f\left(x_i\middle|\theta\right).$$

- A LR test (LRT) is any test that has a rejection of the form $\left\{\mathbf{x} : \lambda\left(\mathbf{x}\right) \leq c\right\}$ where c is any number satisfying $0 \leq c \leq 1$.

- Suppose $\widehat{\theta}$ is the MLE of θ defined by $\widehat{\theta} = \arg\max_{\theta \in \Theta} L\left(\theta \middle| \mathbf{x}\right)$, then we can rewrite

$$\lambda\left(\mathbf{x}\right) = \frac{L\left(\widehat{\theta}_0 \middle| \mathbf{x}\right)}{L\left(\widehat{\theta} \middle| \mathbf{x}\right)}$$

where $\widehat{\theta} = \arg\max_{\theta \in \Theta_0} L\left(\theta \middle| \mathbf{x}\right)$ is the "MLE estimate restricted to $\widehat{\theta}_0$".

- Suppose $\widehat{\theta}$ is the MLE of θ defined by $\widehat{\theta} = \arg\max_{\theta \in \Theta} L\left(\theta \middle| \mathbf{x}\right)$, then we can rewrite

$$\lambda\left(\mathbf{x}\right) = \frac{L\left(\widehat{\theta}_0 \middle| \mathbf{x}\right)}{L\left(\widehat{\theta} \middle| \mathbf{x}\right)}$$

where $\widehat{\theta} = \arg\max_{\theta \in \Theta_0} L\left(\theta \middle| \mathbf{x}\right)$ is the "MLE estimate restricted to $\widehat{\theta}_0$".

- **Example**: Assume we have $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta, 1\right)$ and we want to test $H_0 : \theta = \theta_0$ and $H_1 : \theta \in \Theta_0^c$ then

$$\lambda\left(\mathbf{x}\right) = \frac{L\left(\theta_0 \middle| \mathbf{x}\right)}{L\left(\widehat{\theta} \middle| \mathbf{x}\right)}$$

- It follows that

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\theta_0)^2/2\right)}{(2\pi)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\overline{x})^2/2\right)} \\ &= \exp\left[-n(\overline{x}-\theta_0)^2/2\right].\end{aligned}$$

- It follows that

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\theta_0)^2/2\right)}{(2\pi)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\overline{x})^2/2\right)} \\ &= \exp\left[-n(\overline{x}-\theta_0)^2/2\right].\end{aligned}$$

- An LRT reject H_0 for small values of $\lambda(\mathbf{x})$ and

$$\left\{\mathbf{x}:|\overline{x}-\theta_0|\geq\sqrt{-2(\log c)/n}\right\}$$

where $c\in[0,1]$ thus $\sqrt{-2(\log c)/n}\in[0,\infty)$.

- It follows that

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\theta_0)^2/2\right)}{(2\pi)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\overline{x})^2/2\right)} \\ &= \exp\left[-n(\overline{x}-\theta_0)^2/2\right].\end{aligned}$$

- An LRT reject H_0 for small values of $\lambda(\mathbf{x})$ and

$$\left\{\mathbf{x}:|\overline{x}-\theta_0|\geq\sqrt{-2(\log c)/n}\right\}$$

where $c\in[0,1]$ thus $\sqrt{-2(\log c)/n}\in[0,\infty)$.

- Hence the LRT rejects H_0 is the $\overline{x}$ differs from θ_0 by more than a specified value.

- To construct a size α LRT, we need to select c such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\lambda \left(\mathbf{X} \right) \leq c \right) = \alpha.$$

- To construct a size α LRT, we need to select c such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\lambda \left(\mathbf{X} \right) \leq c \right) = \alpha.$$

- In our case, we have $\Theta_0 = \{\theta_0\}$ so the size α test is

$$\text{Reject } H_0 \text{ if } \left| \overline{x} - \theta_0 \right| \geq \frac{z_{\alpha/2}}{\sqrt{n}}$$

where $P\left(Z \geq z_{\alpha/2} \right) = \frac{\alpha}{2}$ as

$$\begin{aligned} \mathbb{P}_{\theta_0} \left(\left| \overline{X} - \theta_0 \right| \geq \frac{z_{\alpha/2}}{\sqrt{n}} \right) &= \mathbb{P} \left(|Z| > z_{\alpha/2} \right) \text{ where } Z \sim \mathcal{N}(0,1) \\ &= \alpha. \end{aligned}$$

- We have $X_i \overset{\text{i.i.d.}}{\sim} f\left(x \middle| \theta\right)$ with

$$f\left(x \middle| \theta\right) = \exp\left(-x + \theta\right) 1_{[\theta,\infty)}\left(x\right).$$

- We have $X_i \overset{\text{i.i.d.}}{\sim} f\left(x \middle| \theta\right)$ with

$$f\left(x \middle| \theta\right) = \exp\left(-x + \theta\right) 1_{[\theta,\infty)}\left(x\right).$$

- The likelihood is given by

$$L\left(\theta \middle| \mathbf{x}\right) = \exp\left(n\theta - \sum_{i=1}^{n} x_i\right) 1_{\left(-\infty, x_{(1)}\right]}\left(\theta\right).$$

- We have $X_i \overset{\text{i.i.d.}}{\sim} f\left(x|\,\theta\right)$ with

$$f\left(x|\,\theta\right) = \exp\left(-x+\theta\right) 1_{[\theta,\infty)}\left(x\right).$$

- The likelihood is given by

$$L\left(\theta|\,\mathbf{x}\right) = \exp\left(n\theta - \sum_{i=1}^{n} x_i\right) 1_{\left(-\infty, x_{(1)}\right]}\left(\theta\right).$$

- We want to test $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$ and we have $\widehat{\theta}_{MLE} = x_{(1)}$ so

$$\lambda\left(\mathbf{x}\right) = \begin{cases} 1 & \text{if } x_{(1)} \leq \theta_0 \\ \exp\left(-n\left(x_{(1)} - \theta_0\right)\right) & \text{if } x_{(1)} > \theta_0. \end{cases}$$

- We have $X_i \overset{\text{i.i.d.}}{\sim} f(x|\theta)$ with

$$f(x|\theta) = \exp(-x+\theta)\, 1_{[\theta,\infty)}(x).$$

- The likelihood is given by

$$L(\theta|\mathbf{x}) = \exp\left(n\theta - \sum_{i=1}^{n} x_i\right) 1_{\left(-\infty, x_{(1)}\right]}(\theta).$$

- We want to test $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$ and we have $\widehat{\theta}_{MLE} = x_{(1)}$ so

$$\lambda(\mathbf{x}) = \begin{cases} 1 & \text{if } x_{(1)} \leq \theta_0 \\ \exp\left(-n\left(x_{(1)} - \theta_0\right)\right) & \text{if } x_{(1)} > \theta_0. \end{cases}$$

- A LRT test reject H_0 if $\lambda(\mathbf{x}) \leq c$, that is it has the rejection region $\left\{\mathbf{x} : x_{(1)} \geq \theta_0 - \frac{\log c}{n}\right\}$.

- We are looking for a size α test. We have

$$\mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right) = \exp\left(-nc + n\theta_0\right)$$

and

$$\mathbb{P}_{\theta}\left(X_{(1)} \geq c\right) \leq \mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right) \text{ for any } \theta \leq \theta_0.$$

- We are looking for a size α test. We have

$$\mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right) = \exp\left(-nc + n\theta_0\right)$$

and

$$\mathbb{P}_{\theta}\left(X_{(1)} \geq c\right) \leq \mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right) \text{ for any } \theta \leq \theta_0.$$

- Thus we have

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \leq \theta_0} \mathbb{P}_{\theta}\left(X_{(1)} \geq c\right) = \mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right).$$

- We are looking for a size α test. We have

$$\mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right) = \exp\left(-nc + n\theta_0\right)$$

and

$$\mathbb{P}_{\theta}\left(X_{(1)} \geq c\right) \leq \mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right) \text{ for any } \theta \leq \theta_0.$$

- Thus we have

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \leq \theta_0} \mathbb{P}_{\theta}\left(X_{(1)} \geq c\right) = \mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right).$$

- So to get a size α test, we just need to set c such that

$$\mathbb{P}_{\theta_0}\left(X_{(1)} \geq c\right) = \exp\left(-nc + n\theta_0\right) = \alpha.$$

- The LRT is obviously only function of the sufficient statistic which follows from the fact that

$$L\left(\theta|\,\mathbf{x}\right) = h\left(\mathbf{x}\right) g\left(\,T\left(\mathbf{x}\right)|\,\theta\right).$$

- Assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and we want to test $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ where σ^2 is a nuisance parameter.

- Assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and we want to test $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ where σ^2 is a nuisance parameter.
- The LRT can be extended in this case

$$\lambda\left(\mathbf{x}\right) = \frac{\sup_{(\mu,\sigma^2:\mu\leq\mu_0,\sigma^2\geq 0)} L\left(\theta\middle|\,\mathbf{x}\right)}{\sup_{(\mu,\sigma^2:-\infty<\mu<\infty,\sigma^2\geq 0)} L\left(\theta\middle|\,\mathbf{x}\right)} = \frac{\sup_{(\mu,\sigma^2:\mu\leq\mu_0,\sigma^2\geq 0)} L\left(\theta\middle|\,\mathbf{x}\right)}{L\left(\widehat{\theta}_{MLE}\middle|\,\mathbf{x}\right)}$$

- Assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and we want to test $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ where σ^2 is a nuisance parameter.
- The LRT can be extended in this case

$$\lambda\left(\mathbf{x}\right) = \frac{\sup_{(\mu,\sigma^2:\mu\leq\mu_0,\sigma^2\geq 0)} L\left(\theta\middle|\,\mathbf{x}\right)}{\sup_{(\mu,\sigma^2:-\infty<\mu<\infty,\sigma^2\geq 0)} L\left(\theta\middle|\,\mathbf{x}\right)} = \frac{\sup_{(\mu,\sigma^2:\mu\leq\mu_0,\sigma^2\geq 0)} L\left(\theta\middle|\,\mathbf{x}\right)}{L\left(\widehat{\theta}_{MLE}\middle|\,\mathbf{x}\right)}$$

- It follows that

$$\lambda\left(\mathbf{x}\right) = \begin{cases} 1 & \text{if } \widehat{\mu}_{MLE} \leq \mu_0 \\ \frac{L\left(\mu_0,\widehat{\sigma}_0^2\middle|\mathbf{x}\right)L(\mathbf{x})}{L\left(\widehat{\theta}_{MLE}\middle|\mathbf{x}\right)} & \text{if } \widehat{\mu}_{MLE} > \mu_0 \end{cases}$$

as $\left(\mu_0, \widehat{\sigma}_0^2\right)$ is the restricted ML where $\widehat{\sigma}_0^2 = \sum_{i=1}^{n}\left(x_i - \mu_0\right)^2 / n$.

- It would be desirable to **find the test with highest power under H_1 among all size α tests**.

- It would be desirable to **find the test with highest power under H_1 among all size α tests**.
- When such a test exists, it is called **most powerful**.

- It would be desirable to **find the test with highest power under H_1 among all size α tests**.
- When such a test exists, it is called **most powerful**.
- Finding most powerful tests is hard and, in many cases, most powerful tests do not even exist.

- It would be desirable to **find the test with highest power under H_1 among all size α tests**.
- When such a test exists, it is called **most powerful**.
- Finding most powerful tests is hard and, in many cases, most powerful tests do not even exist.
- There is however an important exception when we are testing

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta = \theta_1.$$

- **Theorem** (Neyman-Pearson). Suppose we test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Let

$$T = \frac{\prod_{i=1}^{n} f\left(x_i | \theta_1\right)}{\prod_{i=1}^{n} f\left(x_i | \theta_0\right)}$$

and suppose we reject H_0 when $T > k$. If we choose k so that

$$\mathbb{P}_{\theta_0}\left(T > k\right) = \alpha$$

then **this test is the most powerful size α test**. That is, among all tests with size α, this test maximizes the power $\beta\left(\theta_1\right)$.

- **Theorem** (Neyman-Pearson). Suppose we test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Let

$$T = \frac{\prod_{i=1}^{n} f\left(x_i | \theta_1\right)}{\prod_{i=1}^{n} f\left(x_i | \theta_0\right)}$$

and suppose we reject H_0 when $T > k$. If we choose k so that

$$\mathbb{P}_{\theta_0}\left(T > k\right) = \alpha$$

then **this test is the most powerful size α test**. That is, among all tests with size α, this test maximizes the power $\beta\left(\theta_1\right)$.

- In other words, if the observations is much more likely under H_1 than H_0, we reject it.

- **Example**: Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta, \sigma^2\right)$ where σ^2 is known and we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$.

- **Example**: Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta, \sigma^2\right)$ where σ^2 is known and we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$.
- We have

$$T > k \Leftrightarrow \overline{x} < \frac{\left(2\sigma^2 \log k\right) / n - \theta_0^2 + \theta_1^2}{2\left(\theta_1 - \theta_0\right)}.$$

- **Example**: Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta, \sigma^2\right)$ where σ^2 is known and we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$.
- We have
$$T > k \Leftrightarrow \overline{x} < \frac{\left(2\sigma^2 \log k\right) / n - \theta_0^2 + \theta_1^2}{2\left(\theta_1 - \theta_0\right)}.$$
- The rhs increases from -∞ to ∞ as k goes from 0 to ∞. Thus the test with rejection region $\overline{x} < c$ if the uniformly most powerful level α test, where $\alpha = \mathbb{P}_{\theta_0}\left(\overline{X} < c\right)$.

- **Example**: Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\theta, \sigma^2\right)$ where σ^2 is known and we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$.
- We have
$$T > k \Leftrightarrow \overline{x} < \frac{\left(2\sigma^2 \log k\right) / n - \theta_0^2 + \theta_1^2}{2\left(\theta_1 - \theta_0\right)}.$$
- The rhs increases from -∞ to ∞ as k goes from 0 to ∞. Thus the test with rejection region $\overline{x} < c$ if the uniformly most powerful level α test, where $\alpha = \mathbb{P}_{\theta_0}\left(\overline{X} < c\right)$.
- If a particular α is chosen, then $c = -\sigma z_\alpha / \sqrt{n} + \theta_0$.

- Reporting "reject H_0" or "retain H_0" is not very informative.

- Reporting "reject H_0" or "retain H_0" is not very informative.
- It is also important to report α. If α is small, the decision to reject H_0 is convincing but if α is large it is not because there is a large probability of incorrectly making the decision.

- Reporting "reject H_0" or "retain H_0" is not very informative.
- It is also important to report α. If α is small, the decision to reject H_0 is convincing but if α is large it is not because there is a large probability of incorrectly making the decision.
- An alternative consists of reporting a so-called p-value.

- Reporting "reject H_0" or "retain H_0" is not very informative.
- It is also important to report α. If α is small, the decision to reject H_0 is convincing but if α is large it is not because there is a large probability of incorrectly making the decision.
- An alternative consists of reporting a so-called p-value.
- **Definition**. A p-value $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every $\mathbf{x}$. Small values of $p(\mathbf{x})$ give evidence that H_1 is true. A p-value is valid if for any $0 \leq \alpha \leq 1$

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta (p(\mathbf{X}) \leq \alpha) \leq \alpha.$$

- Reporting "reject H_0" or "retain H_0" is not very informative.
- It is also important to report α. If α is small, the decision to reject H_0 is convincing but if α is large it is not because there is a large probability of incorrectly making the decision.
- An alternative consists of reporting a so-called p-value.
- **Definition**. A p-value $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every $\mathbf{x}$. Small values of $p(\mathbf{x})$ give evidence that H_1 is true. A p-value is valid if for any $0 \leq \alpha \leq 1$

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta (p(\mathbf{X}) \leq \alpha) \leq \alpha.$$

- *Consequence*: If $p(\mathbf{X})$ is a valid p-value then the test that rejects H_0 if and only if $p(\mathbf{X}) \leq \alpha$ is a α level test.

- Reporting "reject H_0" or "retain H_0" is not very informative.
- It is also important to report α. If α is small, the decision to reject H_0 is convincing but if α is large it is not because there is a large probability of incorrectly making the decision.
- An alternative consists of reporting a so-called p-value.
- **Definition**. A p-value $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every $\mathbf{x}$. Small values of $p(\mathbf{x})$ give evidence that H_1 is true. A p-value is valid if for any $0 \leq \alpha \leq 1$

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta (p(\mathbf{X}) \leq \alpha) \leq \alpha.$$

- *Consequence*: If $p(\mathbf{X})$ is a valid p-value then the test that rejects H_0 if and only if $p(\mathbf{X}) \leq \alpha$ is a α level test.
- The smaller the p-value, the stronger the evidence for rejecting H_0.

- Suppose $T(\mathbf{X})$ is a test statistic such that larges values of $T(\mathbf{X})$ give evidence that H_1 is true. For each sample point $\mathbf{x}$, then define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \{ T(\mathbf{X}) \geq T(\mathbf{x}) \},$$

i.e. the *largest possible probability of obtaining a value of* $T(\mathbf{X})$ *which is at least as extreme as the one observed, under the assumption that* H_0 *is true*.

- Suppose $T(\mathbf{X})$ is a test statistic such that larges values of $T(\mathbf{X})$ give evidence that H_1 is true. For each sample point $\mathbf{x}$, then define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \{ T(\mathbf{X}) \geq T(\mathbf{x}) \},$$

  i.e. the *largest possible probability of obtaining a value of* $T(\mathbf{X})$ *which is at least as extreme as the one observed, under the assumption that* H_0 *is true*.
- **Theorem**. $p(\mathbf{x})$ is a valid p-value.

- Suppose $T(\mathbf{X})$ is a test statistic such that larges values of $T(\mathbf{X})$ give evidence that H_1 is true. For each sample point $\mathbf{x}$, then define

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \{ T(\mathbf{X}) \geq T(\mathbf{x}) \},$$

  i.e. the *largest possible probability of obtaining a value of* $T(\mathbf{X})$ *which is at least as extreme as the one observed, under the assumption that* H_0 *is true*.
- **Theorem**. $p(\mathbf{x})$ is a valid p-value.
- **Sketch of Proof**: Let us prove it in the simple case where $\Theta_0 = \{\theta_0\}$. We have in this case

$$\begin{aligned} p(\mathbf{x}) &= \mathbb{P}_{\theta_0} \{ T(\mathbf{X}) \geq T(\mathbf{x}) \} \\ &= \mathbb{P}_{\theta_0} \{ -T(\mathbf{X}) \leq -T(\mathbf{x}) \} \\ &= F_{\theta_0}(-T(\mathbf{x})) \end{aligned}$$

  where F_{θ_0} is the cdf of $-T(\mathbf{X})$. Now we have for any rv Y with cdf F_Y where y is a random variable

$$\begin{aligned} \mathbb{P}_{\theta_0}(F_Y(y) \leq \alpha) &= \mathbb{P}_{\theta_0}\left(y \leq F_Y^{-1}(\alpha)\right) = F_Y\left(F_Y^{-1}(\alpha)\right) \\ &= \alpha. \end{aligned}$$

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"
  - $0.03 < p \leq 0.05$ "Reasonable evidence for rejection"

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"
  - $0.03 < p \leq 0.05$ "Reasonable evidence for rejection"
  - $0.01 < p \leq 0.03$ "Good evidence for rejection"

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"
  - $0.03 < p \leq 0.05$ "Reasonable evidence for rejection"
  - $0.01 < p \leq 0.03$ "Good evidence for rejection"
  - $0.005 < p \leq 0.01$ "Strong evidence for rejection"

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"
  - $0.03 < p \leq 0.05$ "Reasonable evidence for rejection"
  - $0.01 < p \leq 0.03$ "Good evidence for rejection"
  - $0.005 < p \leq 0.01$ "Strong evidence for rejection"
  - $0.001 < p \leq 0.005$ "Very strong evidence for rejection"

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"
  - $0.03 < p \leq 0.05$ "Reasonable evidence for rejection"
  - $0.01 < p \leq 0.03$ "Good evidence for rejection"
  - $0.005 < p \leq 0.01$ "Strong evidence for rejection"
  - $0.001 < p \leq 0.005$ "Very strong evidence for rejection"
  - $0.0005 < p \leq 0.001$ "Extremely strong evidence for rejection"

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"
  - $0.03 < p \leq 0.05$ "Reasonable evidence for rejection"
  - $0.01 < p \leq 0.03$ "Good evidence for rejection"
  - $0.005 < p \leq 0.01$ "Strong evidence for rejection"
  - $0.001 < p \leq 0.005$ "Very strong evidence for rejection"
  - $0.0005 < p \leq 0.001$ "Extremely strong evidence for rejection"
  - $p \leq 0.0005$ "Overwhelming evidence for rejection"

- There is NO magic level which means that null hypotheses are automatically rejected. Although lots of people mistakenly use a significance level of $p < 0.05$ to definitely reject H_0, it should depend entirely on the consequences of being wrong.
- The P-values simply state, under H_0, what the probability that the apparent difference is due to chance.
- At the least you should qualify any statements such as
  - $0.05 < p \leq 0.06$ "Weak evidence for rejection"
  - $0.03 < p \leq 0.05$ "Reasonable evidence for rejection"
  - $0.01 < p \leq 0.03$ "Good evidence for rejection"
  - $0.005 < p \leq 0.01$ "Strong evidence for rejection"
  - $0.001 < p \leq 0.005$ "Very strong evidence for rejection"
  - $0.0005 < p \leq 0.001$ "Extremely strong evidence for rejection"
  - $p \leq 0.0005$ "Overwhelming evidence for rejection"
- P-values are not posterior probabilities!

- **The one sample t-test**: Used for observation $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown parameters.

- **The one sample t-test**: Used for observation $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown parameters.
- We wish to test for location $H_0 : \mu = \mu_0$ or Possible $H_1 : \mu > \mu_0$.

- **The one sample t-test**: Used for observation $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown parameters.
- We wish to test for location $H_0 : \mu = \mu_0$ or Possible $H_1 : \mu > \mu_0$.
- The test statistic is

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu_0\right)}{S} \sim t\left(n-1\right) \text{ (under } H_0\text{)}$$

where S^2 is the sample variance of the X_i's.

- *Example*: We have measured the IQ of 8 students and we wish to see if their intelligence is higher than the population average of 100.

Student	1	2	3	4	5	6	7	8
IQ	118	121	96	102	93	110	117	131

- *Example*: We have measured the IQ of 8 students and we wish to see if their intelligence is higher than the population average of 100.

Student	1	2	3	4	5	6	7	8
IQ	118	121	96	102	93	110	117	131

- $H_0 : \mu = 100$, $H_1 : \mu > 100$. We have $\overline{X} = 111$ and $S^2 = 174$ so $t = 2.36$.

- *Example*: We have measured the IQ of 8 students and we wish to see if their intelligence is higher than the population average of 100.

Student	1	2	3	4	5	6	7	8
IQ	118	121	96	102	93	110	117	131

- $H_0 : \mu = 100$, $H_1 : \mu > 100$. We have $\overline{X} = 111$ and $S^2 = 174$ so $t = 2.36$.
- The p-value is given by

$$P_{H_0}(T > t) = 0.025.$$

- *Example*: We have measured the IQ of 8 students and we wish to see if their intelligence is higher than the population average of 100.

Student	1	2	3	4	5	6	7	8
IQ	118	121	96	102	93	110	117	131

- $H_0 : \mu = 100$, $H_1 : \mu > 100$. We have $\overline{X} = 111$ and $S^2 = 174$ so $t = 2.36$.
- The p-value is given by

$$P_{H_0}(T > t) = 0.025.$$

- Hence there is good evidence for rejection of H_0.

- **The paired t-test**: Suppose we have pairs (X_i, Y_i) and that $D_i = X_i - Y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown parameters.

- **The paired t-test**: Suppose we have pairs (X_i, Y_i) and that $D_i = X_i - Y_i \sim \mathcal{N}(\mu, \sigma^2)$ with unknown parameters.
- We wish to test whether there is a difference in mean between two samples

$$\begin{aligned} H_0 &: \mu = 0 \ (X_i\text{'s don't differ from } Y_i\text{'s}) \\ \text{Possible } H_1 &: \mu \neq \mu_0 \ (X_i\text{'s differ from } Y_i\text{'s}). \end{aligned}$$

- **The paired t-test**: Suppose we have pairs (X_i, Y_i) and that $D_i = X_i - Y_i \sim \mathcal{N}(\mu, \sigma^2)$ with unknown parameters.
- We wish to test whether there is a difference in mean between two samples

$$\begin{aligned} H_0 &: \mu = 0 \ (X_i\text{'s don't differ from } Y_i\text{'s}) \\ \text{Possible } H_1 &: \mu \neq \mu_0 \ (X_i\text{'s differ from } Y_i\text{'s}). \end{aligned}$$

- Test statistics

$$T = \frac{\sqrt{n}\left(\overline{D} - \mu_0\right)}{S_D} \sim t\left(n-1\right) \ \text{(under } H_0\text{)}$$

where S_D^2 is the sample variance of the D_i's.

- *Example*: Two types of rubber (A and B) were randomly assigned to the left and right shoes of 10 children and relative wear on each measure.

i	1	2	3	4	5	6	7	8	9	10
X_i	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
Y_i	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
D_i	-.8	-0.6	-0.3	0.1	-1.1	0.2	-0.3	-0.5	-0.5	-0.3

- *Example*: Two types of rubber (A and B) were randomly assigned to the left and right shoes of 10 children and relative wear on each measure.

i	1	2	3	4	5	6	7	8	9	10
X_i	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
Y_i	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
D_i	-.8	-0.6	-0.3	0.1	-1.1	0.2	-0.3	-0.5	-0.5	-0.3

- We have $\overline{D} = -0.41$, $S_D^2 = 0.15$ so that $t = -3.3489$.

- *Example*: Two types of rubber (A and B) were randomly assigned to the left and right shoes of 10 children and relative wear on each measure.

i	1	2	3	4	5	6	7	8	9	10
X_i	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
Y_i	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6
D_i	-.8	-0.6	-0.3	0.1	-1.1	0.2	-0.3	-0.5	-0.5	-0.3

- We have $\overline{D} = -0.41$, $S_D^2 = 0.15$ so that $t = -3.3489$.
- Noting the form of the alternative hypothesis we calculate

$$P_{H_0}(|T| > t) = 0.0085$$

and so there is strong evidence for rejection of the null hypothesis.

- **The two sample t-test**: Assume we have $X_1, ..., X_m \sim \mathcal{N}\left(\mu_X, \sigma^2\right)$ and $Y_1, ..., Y_n \sim \mathcal{N}\left(\mu_Y, \sigma^2\right)$.

- **The two sample t-test**: Assume we have $X_1, ..., X_m \sim \mathcal{N}\left(\mu_X, \sigma^2\right)$ and $Y_1, ..., Y_n \sim \mathcal{N}\left(\mu_Y, \sigma^2\right)$.
- We wish to test

$$\begin{aligned} H_0 &: \mu_X = \mu_Y \text{ (Null hypothesis)} \\ \text{Possible } H_1 &: \mu_X \neq \mu_Y \text{ (Alternative hypothesis).} \end{aligned}$$

- **The two sample t-test**: Assume we have $X_1, ..., X_m \sim \mathcal{N}\left(\mu_X, \sigma^2\right)$ and $Y_1, ..., Y_n \sim \mathcal{N}\left(\mu_Y, \sigma^2\right)$.
- We wish to test

$$\begin{aligned} H_0 &: \mu_X = \mu_Y \text{ (Null hypothesis)} \\ \text{Possible } H_1 &: \mu_X \neq \mu_Y \text{ (Alternative hypothesis).} \end{aligned}$$

- Use test statistic

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t\left(m + n - 2\right) \text{ (under } H_0\text{)}$$

where the pooled sample variance is

$$S^2 = \frac{\left(m-1\right)S_X^2 + \left(n-1\right)S_Y^2}{m + n - 2}$$

- Under H_0, we have

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right)$$

and also $\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2(m-1)$, $\frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(n-1)$ thus

$$T = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi^2(m+n-2).$$

- Under H_0, we have

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right)$$

and also $\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2(m-1)$, $\frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(n-1)$ thus

$$T = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi^2(m+n-2).$$

- So defining

$$S^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

the result follows.

- *Example*: We have measured the results of two experiments (which we know have the same variance) to determine the concentration of a chemical

Test X	22	19	35	11	21	10
Test Y	33	11	20	38		

- *Example*: We have measured the results of two experiments (which we know have the same variance) to determine the concentration of a chemical

Test X	22	19	35	11	21	10
Test Y	33	11	20	38		

- Test

$$H_0 \quad : \quad \mu_X = \mu_Y \text{ (No difference in mean of experiments)}$$
$$H_1 \quad : \quad \mu_X \neq \mu_Y \text{ (X has a different mean than Y).}$$

- *Example*: We have measured the results of two experiments (which we know have the same variance) to determine the concentration of a chemical

Test X	22	19	35	11	21	10
Test Y	33	11	20	38		

- Test

$$\begin{aligned} H_0 &: \mu_X = \mu_Y \text{ (No difference in mean of experiments)} \\ H_1 &: \mu_X \neq \mu_Y \text{ (X has a different mean than Y).} \end{aligned}$$

- We find $\overline{X} = 19.7$, $\overline{Y} = 25.5$, $S_X = 82.2$ and $S_Y = 90.6$. So $t = -0.87$ and

$$P_{H_0}(|T| > t) = 0.41.$$

Hence we do not reject H_0.

- In the last test, we assume that X and Y have the same variance. How can we check this is the case?

- In the last test, we assume that X and Y have the same variance. How can we check this is the case?
- We use **the F test** which is based on Snecedor's F distribution: If $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ then

$$\frac{U/m}{V/n} \sim F_{m,n}.$$

- In the last test, we assume that X and Y have the same variance. How can we check this is the case?
- We use **the F test** which is based on Snecedor's F distribution: If $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ then

$$\frac{U/m}{V/n} \sim F_{m,n}.$$

- Suppose we have $X_1, ..., X_m \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, ..., Y_n \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and we wish to test

$$\begin{aligned} H_0 &: \sigma_X^2 = \sigma_Y^2 \text{ (Null hypothesis)}, \\ \text{Possible } H_1 &: \sigma_X^2 \neq \sigma_Y^2 \text{ (Alternative hypothesis)}, \end{aligned}$$

then we use the test statistic

$$\frac{S_X^2}{S_Y^2} \sim F_{m-1,n-1} \text{ (under } H_0)$$

- In the last test, we assume that X and Y have the same variance. How can we check this is the case?
- We use **the F test** which is based on Snecedor's F distribution: If $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ then

$$\frac{U/m}{V/n} \sim F_{m,n}.$$

- Suppose we have $X_1, ..., X_m \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, ..., Y_n \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and we wish to test

$$\begin{aligned} H_0 &: \sigma_X^2 = \sigma_Y^2 \text{ (Null hypothesis)}, \\ \text{Possible } H_1 &: \sigma_X^2 \neq \sigma_Y^2 \text{ (Alternative hypothesis)}, \end{aligned}$$

then we use the test statistic

$$\frac{S_X^2}{S_Y^2} \sim F_{m-1,n-1} \text{ (under } H_0)$$

- The greater the ratio deviates from 1, the stronger the evidence for unequal variances.

- Returning to

Test X	22	19	35	11	21	10
Test Y	33	11	20	38		

we find that

$$t = \frac{S_X^2}{S_Y^2} = 0.545$$

and

$$P_{H_0}(|T| > t) = 0.52$$

- Returning to

Test X	22	19	35	11	21	10
Test Y	33	11	20	38		

we find that

$$t = \frac{S_X^2}{S_Y^2} = 0.545$$

and

$$P_{H_0}(|T| > t) = 0.52$$

- It follows that we do not reject the null hypothesis.

- All the tests described so far rely on parametric assumption, we now describe so non-parametric test.

- All the tests described so far rely on parametric assumption, we now describe so non-parametric test.
- Test of location zero: Suppose we have data $D_1, ..., D_n$ and we wish to test H_0 : the data have a symmetric continuous distribution centred about zero.

- All the tests described so far rely on parametric assumption, we now describe so non-parametric test.
- Test of location zero: Suppose we have data $D_1, ..., D_n$ and we wish to test H_0 : the data have a symmetric continuous distribution centred about zero.
- **The sign test** relies on $N_+ =$ Number of data ≥ 0, $N^- =$ Number of data < 0. Under H_0 we have

$$N_+ \sim Bernoulli\,(n, 1/2)$$

and we can easily compute a p-value.

- **Wilconson signed-rank test**: Order the absolute values $|D_1|, ..., |D_n|$ and assign each a rank R_i. The smallest absolute value getting rank 1 and tied scores are assigned a mean rank. If some $D_i = 0$ then we drop these values completely.

- **Wilconson signed-rank test**: Order the absolute values $|D_1|, ..., |D_n|$ and assign each a rank R_i. The smallest absolute value getting rank 1 and tied scores are assigned a mean rank. If some $D_i = 0$ then we drop these values completely.
- Define our statistic to be the sum of the ranks of the positive D_i

$$W^+ = \sum_{D_i > 0} R_i.$$

- **Wilconson signed-rank test**: Order the absolute values $|D_1|, ..., |D_n|$ and assign each a rank R_i. The smallest absolute value getting rank 1 and tied scores are assigned a mean rank. If some $D_i = 0$ then we drop these values completely.
- Define our statistic to be the sum of the ranks of the positive D_i

$$W^+ = \sum_{D_i > 0} R_i.$$

- Extreme values of this statistic (large or small) indicate departure from the null hypothesis. We can work out the exact distribution under H_0 of W^+ using the permutation distribution, otherwise we use a large sample normal approximation.

- *Example*: Suppose we have 9 aptitudes scores of 13,7,3,15,10,12,8,2,9 and we wish to test if they are symmetrically distributed around 10. We first discard the score of 10 leaving 8 scores.

Score	13	7	3	15	12	8	2	9
Score - 10= D	3	-3	-7	5	2	-2	-8	-1
$\|D\|$	3	3	7	5	2	2	8	1
Rank assigned	4.5	4.5	7	6	2.5	2.5	8	1

- *Example*: Suppose we have 9 aptitudes scores of 13,7,3,15,10,12,8,2,9 and we wish to test if they are symmetrically distributed around 10. We first discard the score of 10 leaving 8 scores.

Score	13	7	3	15	12	8	2	9
Score - 10= D	3	-3	-7	5	2	-2	-8	-1
$\lvert D \rvert$	3	3	7	5	2	2	8	1
Rank assigned	4.5	4.5	7	6	2.5	2.5	8	1

- Hence we obtain $w^+ = 13$. After looking at tables, we find that this has a p-value of 0.53 so we do not reject H_0.

- **Multiple testing**. In some situations, we may need to conduct many hypothesis tests.

- **Multiple testing**. In some situations, we may need to conduct many hypothesis tests.
- **Example**. DNA microarrays allow researchers to measure the expression levels of thousands of genes. The data are levels of messenger RNA of each gene. Roughly, the larger the number, the more active the gene. The type of data we have are usually a very large number of genes and two types of patients (say not ill/ill, or ill with disease A/ill with disease B). We have thousands of genes to test.

- **Multiple testing**. In some situations, we may need to conduct many hypothesis tests.
- **Example**. DNA microarrays allow researchers to measure the expression levels of thousands of genes. The data are levels of messenger RNA of each gene. Roughly, the larger the number, the more active the gene. The type of data we have are usually a very large number of genes and two types of patients (say not ill/ill, or ill with disease A/ill with disease B). We have thousands of genes to test.
- Suppose each test is conducted a level α; i.e. the chance of a false rejection of the null is α.

- **Multiple testing**. In some situations, we may need to conduct many hypothesis tests.
- **Example**. DNA microarrays allow researchers to measure the expression levels of thousands of genes. The data are levels of messenger RNA of each gene. Roughly, the larger the number, the more active the gene. The type of data we have are usually a very large number of genes and two types of patients (say not ill/ill, or ill with disease A/ill with disease B). We have thousands of genes to test.
- Suppose each test is conducted a level α; i.e. the chance of a false rejection of the null is α.
- The chance of at least one false rejection is much higher!

- Consider m hypothesis tests

$$H_{0,i} \text{ versus } H_{1,i} \text{ where } i = 1, ..., m$$

- Consider m hypothesis tests

$$H_{0,i} \text{ versus } H_{1,i} \text{ where } i = 1, ..., m$$

- We denote $P_1, ..., P_m$ the m p-values for these tests.

- Consider m hypothesis tests

$$H_{0,i} \text{ versus } H_{1,i} \text{ where } i = 1, ..., m$$

- We denote $P_1, ..., P_m$ the m p-values for these tests.
- To ensure that, the probability of falsely rejecting any null hypothesis is less than or equal to α, we can adopt the Bonferroni method; i.e. reject $H_{0,i}$ is

$$P_i \leq \frac{\alpha}{m}.$$

- **Proof.** Let R be the event that at least one null hypothesis is falsely rejected and let R_i be the event that $H_{0,i}$ is falsely rejected. Then we have

$$\Pr(R) = \Pr\left(\cup_{i=1}^{m} R_i\right) \leq \sum_{i=1}^{m} \Pr(R_i).$$

But by construction we have $\Pr(R_i) = \Pr\left(P_i \leq \frac{\alpha}{m}\right) \leq \frac{\alpha}{m}$ so

$$\Pr(R) \leq \alpha.$$

- **Proof.** Let R be the event that at least one null hypothesis is falsely rejected and let R_i be the event that $H_{0,i}$ is falsely rejected. Then we have

$$\Pr(R) = \Pr\left(\cup_{i=1}^{m} R_i\right) \leq \sum_{i=1}^{m} \Pr(R_i).$$

  But by construction we have $\Pr(R_i) = \Pr\left(P_i \leq \frac{\alpha}{m}\right) \leq \frac{\alpha}{m}$ so

$$\Pr(R) \leq \alpha.$$

- Gene example with $m = 2.638$ genes, we have for $\alpha = .05$ that

$$\frac{\alpha}{m} = 0.0001895375.$$

- **Proof.** Let R be the event that at least one null hypothesis is falsely rejected and let R_i be the event that $H_{0,i}$ is falsely rejected. Then we have

$$\Pr(R) = \Pr\left(\cup_{i=1}^{m} R_i\right) \leq \sum_{i=1}^{m} \Pr(R_i).$$

  But by construction we have $\Pr(R_i) = \Pr\left(P_i \leq \frac{\alpha}{m}\right) \leq \frac{\alpha}{m}$ so

$$\Pr(R) \leq \alpha.$$

- Gene example with $m = 2.638$ genes, we have for $\alpha = .05$ that

$$\frac{\alpha}{m} = 0.0001895375.$$

- The Bonferroni method is far too conservative as it is trying to make it unlikely to have even one single false rejection. So you can expect a lot of type II errors.

- In many scientific problems, the *number of erroneous rejections should be taken into account* and not only the question whether any error was made.

- In many scientific problems, the *number of erroneous rejections should be taken into account* and not only the question whether any error was made.
- The *seriousness of the loss incurred by erroneous rejections is inversely related to the number of hypotheses rejected*.

- In many scientific problems, the *number of erroneous rejections should be taken into account* and not only the question whether any error was made.
- The *seriousness of the loss incurred by erroneous rejections is inversely related to the number of hypotheses rejected*.
- Thus a desirable error rate maybe **the expected proportion of errors among the rejected hypothesis**.

- Suppose we reject all the $H_{0,i}$ such that the p-values fall below a fixed threshold.

- Suppose we reject all the $H_{0,i}$ such that the p-values fall below a fixed threshold.
- Let m_0 be the number of null hypotheses that are true and let $m_1 = m - m_0$.

- Suppose we reject all the $H_{0,i}$ such that the p-values fall below a fixed threshold.
- Let m_0 be the number of null hypotheses that are true and let $m_1 = m - m_0$.
- Then the outcome in multiple testing is

	Retain H_0	Reject H_0	Total
H_0 true	U	V	m_0
H_1 true	T	S	m_1
Total	$m - R$	R	m

- The **False Discovery Proportion** (FDP) is

$$FDP = \begin{cases} V/R = V/(V+S) & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

that is the proportions of rejections (of $H_{0,i}$) that are incorrect.

- The **False Discovery Proportion** (FDP) is

$$FDP = \begin{cases} V/R = V/(V+S) & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

that is the proportions of rejections (of $H_{0,i}$) that are incorrect.

- The **False Discovery Rate** (FDR) is defined as the expectation of the number of false rejections divided by the number of rejections.

$$FDR = \mathbb{E}[FDP]$$

where the expectation is with respect to the unknown distribution of the data.

- The **False Discovery Proportion** (FDP) is

$$FDP = \begin{cases} V/R = V/(V+S) & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

that is the proportions of rejections (of $H_{0,i}$) that are incorrect.

- The **False Discovery Rate** (FDR) is defined as the expectation of the number of false rejections divided by the number of rejections.

$$FDR = \mathbb{E}[FDP]$$

where the expectation is with respect to the unknown distribution of the data.

- Obviously the FDP is NOT observed and the FDR cannot be computed either!

- If all null hypothesis are true, i.e. $m_0 = m$, then $S = 0$ and $FDP = 1$ if $V \geq 1$ and 0 otherwise; so

$$FDR = \mathbb{E}\left[FDP\right] = \Pr\left(V \geq 1\right)$$

and the FDR is just the probability of committing any type I error (the probability "controlled" by Bonferroni). Note that in this case, there cannot be any type II error as $m_1 = 0$.

- If all null hypothesis are true, i.e. $m_0 = m$, then $S = 0$ and $FDP = 1$ if $V \geq 1$ and 0 otherwise; so

$$FDR = \mathbb{E}\left[FDP\right] = \Pr\left(V \geq 1\right)$$

and the FDR is just the probability of committing any type I error (the probability "controlled" by Bonferroni). Note that in this case, there cannot be any type II error as $m_1 = 0$.

- When $m_0 < m$, then we have if $V \geq 1$ then

$$V/R = V/\left(V + S\right) \leq 1$$

thus $\mathbb{I}_{\{V \geq 1\}} \geq V/R$ thus

$$FDR \leq \Pr\left(V \geq 1\right).$$

- If all null hypothesis are true, i.e. $m_0 = m$, then $S = 0$ and $FDP = 1$ if $V \geq 1$ and 0 otherwise; so

$$FDR = \mathbb{E}\left[FDP\right] = \Pr\left(V \geq 1\right)$$

and the FDR is just the probability of committing any type I error (the probability "controlled" by Bonferroni). Note that in this case, there cannot be any type II error as $m_1 = 0$.

- When $m_0 < m$, then we have if $V \geq 1$ then

$$V/R = V/\left(V + S\right) \leq 1$$

thus $\mathbb{I}_{\{V \geq 1\}} \geq V/R$ thus

$$FDR \leq \Pr\left(V \geq 1\right).$$

- The larger the number of non-true null hypothesis is, the larger S tends to be and the difference between FDR and $\Pr\left(V \geq 1\right)$ increase.

- If all null hypothesis are true, i.e. $m_0 = m$, then $S = 0$ and $FDP = 1$ if $V \geq 1$ and 0 otherwise; so

$$FDR = \mathbb{E}\left[FDP\right] = \Pr\left(V \geq 1\right)$$

  and the FDR is just the probability of committing any type I error (the probability "controlled" by Bonferroni). Note that in this case, there cannot be any type II error as $m_1 = 0$.
- When $m_0 < m$, then we have if $V \geq 1$ then

$$V/R = V/\left(V + S\right) \leq 1$$

  thus $\mathbb{I}_{\{V \geq 1\}} \geq V/R$ thus

$$FDR \leq \Pr\left(V \geq 1\right).$$

- The larger the number of non-true null hypothesis is, the larger S tends to be and the difference between FDR and $\Pr\left(V \geq 1\right)$ increase.
- The advantage of FDR is that if you can control it, then you can expect to increase the power and limit type II errors.

- The Benjamini-Hochberg method (1995) allows us to control the FDR.

- The Benjamini-Hochberg method (1995) allows us to control the FDR.

1. Let $P_{(1)} < \cdots < P_{(m)}$ denoted the ordered independent P-values.

- The Benjamini-Hochberg method (1995) allows us to control the FDR.

1. Let $P_{(1)} < \cdots < P_{(m)}$ denoted the ordered independent P-values.
2. Define

$$l_i = \frac{i.\alpha}{m} \text{ and } R = \max\left\{i : P_{(i)} < l_i\right\}$$

- The Benjamini-Hochberg method (1995) allows us to control the FDR.

1. Let $P_{(1)} < \cdots < P_{(m)}$ denoted the ordered independent P-values.
2. Define
$$l_i = \frac{i.\alpha}{m} \text{ and } R = \max\left\{i : P_{(i)} < l_i\right\}$$
3. Let $T = P_{(R)}$ be the so-called BH rejection treshold.

- The Benjamini-Hochberg method (1995) allows us to control the FDR.

1. Let $P_{(1)} < \cdots < P_{(m)}$ denoted the ordered independent P-values.
2. Define
$$l_i = \frac{i.\alpha}{m} \text{ and } R = \max\left\{i : P_{(i)} < l_i\right\}$$
3. Let $T = P_{(R)}$ be the so-called BH rejection treshold.
4. Reject all $H_{0,i}$ for which $P_i \leq T$.

- **Theorem.** Regardless of how many nulls are true and regardless of the distribution of the p-values when the null hypothesis is false,

$$FDR = \mathbb{E}\left(FDP\right) \leq \frac{m_0}{m}\alpha \leq \alpha.$$

- **Theorem.** Regardless of how many nulls are true and regardless of the distribution of the p-values when the null hypothesis is false,

$$FDR = \mathbb{E}\left(FDP\right) \leq \frac{m_0}{m}\alpha \leq \alpha.$$

- Proof: one full paper JRSSB... not presented here.

- **Theorem.** Regardless of how many nulls are true and regardless of the distribution of the p-values when the null hypothesis is false,

$$FDR = \mathbb{E}\left(FDP\right) \leq \frac{m_0}{m}\alpha \leq \alpha.$$

- Proof: one full paper JRSSB... not presented here.
- It is shown that the 'average power' (i.e. the probability of H_1 correctly rejected) is much higher than for the Bonferroni test.

- Suppose 10 independent hypothesis tests are carried leading to the following ordered p-values

i	1	2	3	4	5
$p_{(i)}$	0.00017	0.00448	0.00671	0.00907	0.01220

i	6	7	8	9	10
$p_{(i)}$	0.3362	0.39341	0.53882	0.58125	0.98617

- Suppose 10 independent hypothesis tests are carried leading to the following ordered p-values

i	1	2	3	4	5
$p_{(i)}$	0.00017	0.00448	0.00671	0.00907	0.01220

i	6	7	8	9	10
$p_{(i)}$	0.3362	0.39341	0.53882	0.58125	0.98617

- When $\alpha = 0.05$, the Bonferroni test rejects any p-value less than $\alpha/10 = 0.005$. Thus, only the first two hypotheses are rejected.

- Suppose 10 independent hypothesis tests are carried leading to the following ordered p-values

i	1	2	3	4	5
$p_{(i)}$	0.00017	0.00448	0.00671	0.00907	0.01220

i	6	7	8	9	10
$p_{(i)}$	0.3362	0.39341	0.53882	0.58125	0.98617

- When $\alpha = 0.05$, the Bonferroni test rejects any p-value less than $\alpha/10 = 0.005$. Thus, only the first two hypotheses are rejected.
- For the BH test, we find the largets i such that $P_{(i)} < i\alpha/m = 0.0005i$ which in this case is $i = 5$. Thus we reject the first five hypotheses.