50.034 - Introduction to Probability and Statistics

Week 13 – Lecture 24 (Review Lecture)

January–May Term, 2019

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Lecture 24 Assigned Readings: (Revise: All topics covered in this course.)

Outline of Lecture

- Statistical model, parameter space, statistic
- Prior and posterior distributions
- Families of conjugate priors
- Estimators and Estimates
- Special distributions (beta, gamma, χ^2, t)
- Useful results on χ^2 distribution and t-distribution
- Confidence intervals and hypothesis testing
- Error, significance level, power, p-value
- Likelihood statistic, likelihood ratio test, likelihood ratio
- t-test, two-sample t-test
- χ^2 test of goodness of fit

Due to the lack of time, we shall review the least squares method and linear regression during cohort class this week.

Statistical model, parameter space, statistic

Definition: A statistical model consists of the following:

- A collection of R.V.'s $\{X_1, X_2, X_3, \dots\}$ (could be finite or infinite)
  - These R.V.'s could be observable or latent.
- A family of possible joint distributions for observable R.V.'s.
- Assumptions on the parameters of the joint distributions.
  - e.g. parameter λ is a R.V. with uniform distribution

Definition: The parameters of a distribution are numerical attributes whose values determine the distribution completely.

- e.g. binomial distribution with parameters n and p.
- Given any parameter θ, the set of all possible values for θ is called the parameter space of θ.
  - What is considered "possible" depends on the context.

Definition: Let $\mathcal{S} = \{X_1, \dots, X_n\}$ be a set of n observable R.V.'s. A statistic of $\mathcal{S}$ is a function of the R.V.'s in $\mathcal{S}$.

- **Note:** A statistic is a random variable!
- **Interpretation:** A statistic is a descriptive summary of some given set of observable R.V.'s.

Prior and posterior distributions

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$. Let θ be a parameter (possibly one of many parameters) of the joint distribution of $X_1, \ldots, X_n$, and treat θ as a random variable.

The prior distribution of θ is the initial distribution specified for θ.

- This is the distribution we specify before observing any data (i.e. before gathering the observed values for $X_1, \ldots, X_n$)
- "prior distribution" can simply be called "prior".

After we have some observed values, say $X_1 = x_1, \ldots, X_n = x_n$, then the conditional distribution, consisting of all conditional probabilities of the form $\Pr(\theta \in C | X_1 = x_1, \ldots, X_n = x_n)$ (over all possible $C \subseteq \mathbb{R}$), is called the posterior distribution of θ.

- "posterior distribution" can simply be called "posterior".

Interpretation: The prior of θ is the initial guess for the distribution of θ, while the posterior of θ is the updated guess, after taking into account the observed values $X_1 = x_1, \ldots, X_n = x_n$.

Prior pmf/pdf versus Posterior pmf/pdf

Consider a statistical model with observable R.V.'s $X_1, \ldots, X_n$.
Suppose θ is a parameter of the joint distribution of $X_1, \ldots, X_n$, where θ is treated as a random variable.

- If θ is discrete, then the pmf of θ is called the prior pmf of θ.
- If θ is continuous, then the pdf of θ is called the prior pdf of θ.
- In either case, the pmf/pdf of θ is usually written as $\xi(\theta)$.

Next, suppose we have observed the values $X_1 = x_1, \ldots, X_n = x_n$.

- If θ is discrete, then the posterior pmf of θ is the conditional pmf of θ given $X_1 = x_1, \ldots, X_n = x_n$.
- If θ is continuous, then the posterior pdf of θ is the conditional pdf of θ given $X_1 = x_1, \ldots, X_n = x_n$.
- In either case, the pmf/pdf is denoted by $\xi(\theta | x_1, \ldots, x_n)$, or more simply, $\xi(\theta|\mathbf{x})$, where $\mathbf{x}$ represents $(x_1, \ldots, x_n)$.

$$\boxed{\text{posterior distribution}} = \boxed{\text{conditional distribution of the parameter given the data/evidence}}$$

Families of conjugate priors

Let Ψ be a family of distributions.

Definition: Consider a statistical model where $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ. We say that Ψ is a conjugate family of prior distributions, or more simply, a family of conjugate priors, if the following condition holds:

- If the prior distribution of θ is chosen from Ψ, then the posterior distribution of θ will also be in Ψ, no matter what n is, and not matter what the observed values of $X_1, \ldots, X_n$ are.

Examples of families of conjugate priors:

Sampling from	Family of conjugate priors
Bernoulli distribution	beta distributions
binomial distribution*	beta distributions
geometric distribution	beta distributions
Poisson distribution	gamma distributions
exponential distribution	gamma distributions
normal distribution	normal distributions

*Note: For a fixed known number of trials.

Sampling from various distributions

Consider a statistical model where $X_1, \dots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

Theorem: (Sampling from Bernoulli distributions)
If $X_1, \dots, X_n$ are Bernoulli, and if θ has the beta prior distribution with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \dots, X_n = x_n$ is the beta distribution with parameters $\alpha + (x_1 + \cdots + x_n)$ and $\beta + n - (x_1 + \cdots + x_n)$.

- i.e. the new parameters of the beta posterior are $\alpha' = \alpha+(\text{number of successes}), \quad \beta' = \beta+(\text{number of failures})$.

Theorem: (Sampling from binomial distributions)
Let $N \geq 1$ be a known fixed integer. If $X_1, \dots, X_n$ are binomial with parameters N and θ, and if θ has the beta prior with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \dots, X_n = x_n$ is the beta distribution with parameters $\alpha + (x_1 + \cdots + x_n)$ and $\beta + Nn - (x_1 + \cdots + x_n)$.

- i.e. the new parameters of the beta posterior are
$$\alpha' = \alpha + \begin{pmatrix}\text{total number of}\\ \text{successes}\end{pmatrix}, \quad \beta' = \beta + \begin{pmatrix}\text{total number of}\\ \text{failures}\end{pmatrix}.$$

Sampling from various distributions (continued)

Consider a statistical model where $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

Theorem: (Sampling from Poisson distributions)
If $X_1, \ldots, X_n$ are Poisson, and if θ has the gamma prior distribution with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$ is the gamma distribution with parameters $\alpha + (x_1 + \cdots + x_n)$ and $\beta + n$.

- i.e. the new parameters of the gamma posterior are

$$\alpha' = \alpha + \binom{\text{number of new}}{\text{occurrences}}, \qquad \beta' = \beta + \binom{\text{number of time}}{\text{periods}}.$$

Theorem: (Sampling from exponential distributions)
If $X_1, \ldots, X_n$ are exponential, and if θ has the gamma prior with parameters α and β, then the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$ is the gamma distribution with parameters $\alpha + n$ and $\beta + (x_1 + \cdots + x_n)$.

- i.e. the new parameters of the gamma posterior are

$$\alpha' = \alpha + \binom{\text{number of}}{\text{experiments}}, \qquad \beta' = \beta + \binom{\text{total time}}{\text{elapsed}}.$$

Sampling from various distributions (continued)

Consider a statistical model where $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

Theorem: (**Sampling from normal distributions**)
Let $\sigma > 0$ be a fixed known real number. If $X_1, \ldots, X_n$ are normal with mean θ and variance σ^2, and if θ has the normal prior distribution with mean μ_0 and variance v_0^2, then the posterior distribution of θ given $X_1 = x_1, \ldots, X_n = x_n$ is the normal distribution with mean μ_1 and variance v_1^2 given as follows:

$$
\begin{aligned}
\mu_1 &= \frac{\sigma^2 \mu_0 + v_0^2 (x_1 + \cdots + x_n)}{\sigma^2 + n v_0^2}, \\
v_1^2 &= \frac{\sigma^2 v_0^2}{\sigma^2 + n v_0^2}.
\end{aligned}
$$

Estimators and Estimates

Let $X_1, \dots, X_n$ be observable R.V.'s whose joint distribution is parametrized by a parameter θ.

- An estimator of θ is a real-valued function $\delta(X_1, \dots, X_n)$.
- Given δ and a vector $\mathbf{x} = (x_1, \dots, x_n)$ of observed values, the real number $\delta(\mathbf{x})$ is called an estimate of θ.

Note: An estimator is a statistic.

- **Recall:** A statistic is a function of observable R.V.'s.

Examples of estimators:

- (Lecture 15) Bayes estimator $\delta^*(X_1, \dots, X_n)$
  - δ^* minimizes Bayes risk over all possible estimates.
  - Given an estimate $a = \delta^*(\mathbf{x})$ and a loss function $L(x, y)$, the Bayes risk of δ^* is the expected loss $\mathbf{E}[L(\theta, a)|\mathbf{x}]$.
- (Lecture 16) Maximum likelihood estimator $\hat{\theta}(X_1, \dots, X_n)$
  - $\hat{\theta}$ maximizes likelihood function over all possible estimates.
  - The likelihood function of θ is defined using the exact same expression for the joint condition pmf/pdf (either $p_n(\mathbf{x}|\theta)$ or $f_n(\mathbf{x}|\theta)$), but treated as a function only in terms of θ.

Posterior mean as an estimator

Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by a parameter θ.

Theorem: (Lecture 15) The **Bayes estimator** of θ with respect to the **squared error loss function** $L(x, y) = (x - y)^2$ is the estimator

$$\delta^*(X_1, \ldots, X_n) = \mathbf{E}[\theta | X_1, \ldots, X_n].$$

- **Definition:** $\mathbf{E}[\theta | X_1, \ldots, X_n]$, treated as a estimator, is called the posterior mean of θ.
- **Note:** $\mathbf{E}[\theta | X_1, \ldots, X_n]$ is a function of $X_1, \ldots, X_n$, similar to how we saw in Lecture 9 that $\mathbf{E}[X|Y]$ is a function of Y.

Technicality: Frequently, the posterior mean is treated as an estimator (i.e. a function of $X_1, \ldots, X_n$). However, the posterior mean is also sometimes treated as an estimate, i.e. a real number $\mathbf{E}[\theta | x_1, \ldots, x_n]$ given some actual observed values $x_1, \ldots, x_n$. Whether the posterior mean is a function or a real number will depend on the context. For example, if the observed values are not given, then it is implicitly assumed that the posterior mean is treated as an estimator.

Unbiased versus biased estimators

Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by some parameter θ with parameter space Ω.
Let $\delta = \delta(X_1, \ldots, X_n)$ be an estimator of $\{X_1, \ldots, X_n\}$.

- **Recall:** The sampling distribution of δ is the distribution of δ.
- For every possible value θ in Ω, the mean of the sampling distribution of δ given $\theta = \theta$, is denoted by $\mathbf{E}_\theta[\delta(X_1, \ldots, X_n)]$.

Definition: We say that δ is unbiased if $\mathbf{E}_\theta[\delta(X_1, \ldots, X_n)] = \theta$ for every possible value θ in Ω, and we say that δ is biased otherwise.

- The bias of δ is a function defined on Ω, such that each $\theta \in \Omega$ is mapped to $\mathbf{E}_\theta[\delta(X_1, \ldots, X_n)] - \theta$.

Interpretation: Let $\delta = \delta(X_1, \ldots, X_n)$ be an estimator of some parameter θ with parameter space Ω. If for every possible value θ in Ω, the mean of the estimator is exactly θ, then the bias of δ is the zero function.

Biased versus unbiased sample variance

Let $\{X_1, \ldots, X_n\}$ be a random sample with mean μ and variance σ^2.

- The biased sample variance of $\{X_1, \ldots, X_n\}$ is

$$\hat{\sigma}_n^2 = \hat{\sigma}_n^2(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

- The unbiased sample variance of $\{X_1, \ldots, X_n\}$ is

$$s_n^2 = s_n^2(X_1, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Facts: $\mathbf{E}[s_n^2] = \sigma^2$ (for $n > 1$), while $\mathbf{E}[\hat{\sigma}_n^2] = \frac{n-1}{n}\sigma^2$ (for all $n \geq 1$).

- The biased sample variance has negative bias $-\frac{\sigma^2}{n}$.
  - $\hat{\sigma}_n^2$ consistently underestimates the "true" variance.
  - This negative bias approaches 0 as $n \to \infty$.
- The unbiased sample variance has zero bias.

Theorem: If $X_1, \ldots, X_n$ are **normal** R.V.'s, then the **maximum likelihood estimator** of σ^2 is the biased sample variance $\hat{\sigma}_n^2$.

Special distributions

Beta distribution (with parameters α and β)

- **Common Use:** Prior/posterior distribution of the success rate of a **Bernoulli process**.
- Special case: uniform distribution ($\alpha = 1, \beta = 1$).

Gamma distribution (with parameters α and β)

- **Common Uses:** Prior/posterior distribution of the parameter of a **Poisson process** or an **exponential process**.
- Special case: exponential distribution with parameter β ($\alpha = 1$).

Chi-squared distribution (with *m* degrees of freedom)

- **Common Uses:** To model the **sample variances** of random samples of normal R.V.'s; to model the **sum of squares** of standard normal R.V.'s; to model goodness of fit.
- Special case: exponential distribution with parameter $\frac{1}{2}$ ($m = 2$).

t-distribution (with *m* degrees of freedom)

- **Common Uses:** Uses analogous to the standard normal distribution in the case of **small sample size** and/or **unknown variance**.

Useful results involving χ^2 distribution

Theorem: If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$.

Theorem: Let $Y_1, \dots, Y_n$ be **independent** R.V.'s, such that $Y_i \sim \chi^2(m_i)$ for each $1 \le i \le m$. Then the sum $Y_1 + \cdots + Y_n$ has the χ^2 distribution with $m_1 + \cdots + m_n$ degrees of freedom.

Corollary: Let $Z_1, \dots, Z_n$ be iid **standard normal** R.V.'s. Then $(Z_1^2 + \cdots + Z_n^2) \sim \chi^2(n)$.

Corollary: Let $X_1, \dots, X_n$ be iid **normal** R.V.'s with mean μ and variance σ^2. Then $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2(n)$.

Theorem: Let $\{X_1, \dots, X_n\}$ be a random sample of observable **normal** R.V.'s with variance σ^2, biased sample variance $\hat{\sigma}^2$, and sample mean $\hat{\mu}$. Then $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \hat{\mu})^2 = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$.

Useful results involving t-distributions

Theorem: Let $\{X_1, \ldots, X_n\}$ be a random sample of **normal** R.V.'s with mean μ and variance σ^2. Let $\overline{X}_n$ and s_n^2 be the sample mean and the **unbiased sample variance** respectively. Then $\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}$ has the t-distribution with $(n-1)$ degrees of freedom.

- In comparison, $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ has the standard normal distribution.

Theorem: For each $n \geq 1$, let Z_n be the R.V. that has the t-distribution with n degrees of freedom. Then the asymptotic distribution of the infinite sequence $Z_1, Z_2, Z_3, \ldots$ is the standard normal distribution.

Intuition: As $n \to \infty$, the unbiased sample variance s_n^2 approaches the "true" variance σ^2, so $\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}$ would become approximately $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. Therefore, for a sufficiently large degree of freedom, the t-distribution is approximately standard normal.

Confidence Intervals of Parameters

Let $0 < p < 1$, and let $\{X_1, \ldots, X_n\}$ be a random sample of observable R.V.'s that depend on some parameter θ.

- If T_1 and T_2 are statistics such that $\Pr(T_1 < \theta < T_2) \geq p$ for all possible values of θ, then we say that the random open interval (T_1, T_2) is a $100p$ percent confidence interval for θ.
    - We say that the confidence level is $100p$ percent.
    - If $\Pr(T_1 < \theta < T_2) = p$ for all possible values of θ, then the confidence interval (T_1, T_2) is called exact.
- Given observed values $T_1 = t_1$ and $T_2 = t_2$, the open interval (t_1, t_2) is called the observed value of the confidence interval.

Important Note: A confidence interval is a **pair of statistics** forming a random open interval.

Interpretation: By saying that (T_1, T_2) is a 95% confidence interval for θ, it means that 95% of all observed values (t_1, t_2) for (T_1, T_2) are open intervals that actually contain θ.

- It does **NOT** mean every observed open interval (t_1, t_2) has a 95% probability of containing θ. The "95%" relates to the entire estimation procedure, and not to a specific open interval.

Hypothesis Testing

Goal: Perform hypothesis testing on the parameter θ.

1. Specify some **null hypothesis** $H_0 : \theta \in \Omega_0$.
   - $\Omega_0 \subseteq \Omega$ is a subset chosen based on your specific application.
   - You wish to test whether the "true" value of θ is not in Ω_0.
2. Specify some **test statistic** $T = T(X_1, \ldots, X_n)$.
   - Your final decision will depend on the observed value of T.
3. Specify some **rejection region** $R \subseteq \mathbb{R}$.
   - This represents the region for where to reject H_0.
   - Note: R can be different from the complement of Ω_0.
4. Collect experimental evidence
   - Get observed values $X_1 = x_1, \ldots, X_n = x_n$.
5. Final decision: To reject or not to reject?
   - "Reject H_0" if $T(x_1, \ldots, x_n) \in R$.
   - "Do not reject H_0" if $T(x_1, \ldots, x_n) \notin R$.

The entire test procedure is collectively called a hypothesis test.

- A type I error occurs if H_0 is **true** but we **reject** H_0.
- A type II error occurs if H_0 is **false** but we **do not reject** H_0.

Errors, significance level, power

Let $\mathcal{H}$ be a hypothesis test with null hypothesis $H_0 : \theta \in \Omega_0$.
Let Ω be the parameter space of θ, and let $\Omega_1 = \Omega \backslash \Omega_0$.
Let T be the test statistic, and let R be the rejection region.

Definition: The power function of $\mathcal{H}$ is $\pi(\omega) = \Pr(T \in R | \theta = \omega)$, defined for every possible value $\omega \in \Omega$.

- **Interpretation:** $\pi(\omega)$ is the probability that we will **reject** the null hypothesis H_0, given that the "true" value of θ equals ω.

Definition: We say that $\mathcal{H}$ a level α_0 test, or equivalently, that $\mathcal{H}$ has a significance level of α_0, if $\pi(\omega) \leq \alpha_0$ for all $\omega \in \Omega_0$.

- **Interpretation:** "$\mathcal{H}$ is a level α_0 test" is exactly the same as "the probability that a type I error occurs for $\mathcal{H}$ is at most α_0."
  - The smallest possible α_0 is called the size of $\mathcal{H}$.

Definition: Let β_0 be a real number. We say that $\mathcal{H}$ has a power of β_0, if $\pi(\omega) \geq \beta_0$ for all $\omega \in \Omega_1$.

- **Interpretation:** "$\mathcal{H}$ has power β_0" is exactly the same as "the probability that a type II error occurs is at most $1 - \beta_0$".
  - Higher power implies lower probability that type II errors occur.

p-value

Let $T = T(X_1, \ldots, X_n)$ be a fixed statistic of a random sample $\{X_1, \ldots, X_n\}$ of observable R.V.'s with unknown parameter θ.

Let $\mathcal{H} = \{\mathcal{H}_c\}_{c \in \mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represents the hypothesis test with null hypothesis $H_0 : \theta \in \Omega_0$, test statistic T, and rejection region $[c, \infty)$. [Note: Ω_0 is a fixed subset.]

- Let α_c be the **size** of each $\mathcal{H}_c$, i.e. α_c is the smallest possible significance level for $\mathcal{H}_c$. (Different values of c give different sizes.)

Definition: Given some observed values $X_1 = x_1, \ldots, X_n = x_n$, let $t = T(x_1, \ldots, x_n)$ be the corresponding observed value of T. Then as c varies over $\mathbb{R}$, the smallest possible size α_c for which H_0 will be rejected given the observed value t, is called the p-value of $\mathcal{H}$.

- **Note:** The p-value depends on the observed values $x_1, \ldots, x_n$.
- **Interpretation:** If α is the p-value of $\mathcal{H}$, then it means that our experimental data is sufficient evidence to reject the null hypothesis H_0, whenever the value of c is chosen such that $\mathcal{H}_c$ has a significance level $\alpha_c \geq \alpha$.

Likelihood ratio statistic and likelihood ratio test

Let θ be the parameter of some observable R.V.'s $X_1, \ldots, X_n$, and let Ω be the parameter space of θ.

- Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be the M.L.E. of θ. By definition, $\hat{\theta}$ maps each possible vector of observed values $\mathbf{x} = (x_1, \ldots, x_n)$ for $(X_1, \ldots, X_n)$ to some value in Ω that maximizes the likelihood function of θ (either $p_n(\mathbf{x}|\theta)$ or $f_n(\mathbf{x}|\theta)$), thus

$$p_n(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \sup_{\theta\in\Omega} p_n(\mathbf{x}|\theta) \qquad \text{or} \qquad f_n(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \sup_{\theta\in\Omega} f_n(\mathbf{x}|\theta).$$

Definition: Given a subset $\Omega_0 \subseteq \Omega$, the likelihood ratio statistic associated to Ω_0 is the statistic $\Lambda = \Lambda(X_1, \ldots, X_n)$ defined by

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Omega_0} p_n(\mathbf{x}|\theta)}{\sup_{\theta\in\Omega} p_n(\mathbf{x}|\theta)} \qquad \text{or} \qquad \Lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Omega_0} f_n(\mathbf{x}|\theta)}{\sup_{\theta\in\Omega} f_n(\mathbf{x}|\theta)}$$

for each possible vector of observed values $\mathbf{x} = (x_1, \ldots, x_n)$.

Definition: If $\mathcal{H}$ has the **likelihood ratio statistic** associated to Ω_0 as its test statistic, then we say that $\mathcal{H}$ is a likelihood ratio test.

Likelihood ratio and Neyman–Pearson lemma

Let $\mathcal{H}$ be a hypothesis test with the following **simple** hypotheses:

- null hypothesis $H_0 : \theta = \theta_0$; alternative hypothesis $H_1 : \theta = \theta_1$;

where θ is a random vector of parameters for the joint distribution of $X_1, \ldots, X_n$. Let $\mathcal{L}(\theta|\mathbf{x})$ be the likelihood function of θ (given $\mathbf{x}$).

Theorem: Let $a, b > 0$ be constants. Suppose the test statistic Λ of $\mathcal{H}$ is defined by $\Lambda(\mathbf{x}) = \frac{\mathcal{L}(\theta_1|\mathbf{x})}{\mathcal{L}(\theta_0|\mathbf{x})}$ for each possible $\mathbf{x}$, and let the rejection region of $\mathcal{H}$ be $(\frac{a}{b}, \infty)$ or $[\frac{a}{b}, \infty)$. Then every test $\mathcal{H}'$ with the same H_0, H_1 satisfies $a\alpha(\mathcal{H}) + b\beta(\mathcal{H}) \leq a\alpha(\mathcal{H}') + b\beta(\mathcal{H}')$.

- The ratio $\frac{\mathcal{L}(\theta_1|\mathbf{x})}{\mathcal{L}(\theta_0|\mathbf{x})}$ is called the likelihood ratio of $\mathbf{x}$.
- **Note:** The likelihood ratio $\neq$ likelihood ratio statistic!

Neyman–Pearson lemma: If $\mathcal{H}'$ is another test with the same hypotheses H_0 and H_1, but with a smaller type I error probability, i.e. $\alpha(\mathcal{H}') < \alpha(\mathcal{H})$, then its type II error probability must be larger, i.e. $\beta(\mathcal{H}') > \beta(\mathcal{H})$, or equivalently, $\mathcal{H}'$ must have a smaller **power**.

- $\mathcal{H}$ is called most powerful at significance level α_0 if the power of $\mathcal{H}$ is maximum among all level α_0 tests. (with the same hypotheses)

t-test

Definition: A t-test is any hypothesis test with null hypothesis $H_0 : \theta \in \Omega_0$, such that the test statistic has the **t-distribution** for some specific $\theta = \theta_0$ in Ω_0.

Three most important examples of t-tests:
Let $\{X_1, \ldots, X_n\}$ be a random sample of **normal** observable R.V.'s with **unknown mean** μ and **unknown variance** σ^2. Let $\overline{X}_n$, s_n^2 be the sample mean and the **unbiased sample variance** respectively. Let μ_0 be some real constant, and define the R.V. $T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{s_n}$.

- If $\mathcal{H}$ is a hypothesis test with $H_0 : \mu \leq \mu_0$, test statistic T, and rejection region $[c, \infty)$, then $\mathcal{H}$ is a one-sided t-test.
- If $\mathcal{H}$ is a hypothesis test with $H_0 : \mu \geq \mu_0$, test statistic T, and rejection region $(-\infty, c]$, then $\mathcal{H}$ is a one-sided t-test.
- If $\mathcal{H}$ is a hypothesis test with $H_0 : \mu = \mu_0$, test statistic $|T|$, and rejection region $[c, \infty)$, then $\mathcal{H}$ is a two-sided t-test.

Significance levels and p-values of t-tests

Let $\{X_1, \ldots, X_n\}$ be a random sample of **normal** observable R.V.'s with **unknown mean** μ and **unknown variance** σ^2. Let $\overline{X}_n$, s_n^2 be the sample mean and the **unbiased sample variance** respectively. Let μ_0 and c_0 be fixed real numbers, and define $T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{s_n}$.

Theorem: Suppose that c_0 is the $100(1-\alpha_0)$-percentile of the t-distribution with $n-1$ degrees of freedom.

1. Let $\mathcal{H}$ be a t-test with null hypothesis $H_0 : \mu \leq \mu_0$, test statistic T, and rejection region $[c, \infty)$.
   - $\mathcal{H}$ has significance level α_0 if and only if $c \geq c_0$.
   - If the observed value of T is c_0, then the p-value of $\mathcal{H}$ is α_0.
2. Let $\mathcal{H}$ be a t-test with null hypothesis $H_0 : \mu \geq \mu_0$, test statistic T, and rejection region $(-\infty, c]$.
   - $\mathcal{H}$ has significance level α_0 if and only if $c \leq c_0$.
   - If the observed value of T is c_0, then the p-value of $\mathcal{H}$ is α_0.
3. Let $\mathcal{H}$ be the t-test with null hypothesis $H_0 : \mu = \mu_0$, test statistic $|T|$, and rejection region $[c, \infty)$.
   - $\mathcal{H}$ has significance level $2\alpha_0$ if and only if $c \geq c_0$.
   - If the observed value of $|T|$ is c_0, then the p-value of $\mathcal{H}$ is $2\alpha_0$.

Two-sample t-statistic

Note: t-tests also make sense on two random samples.

- Let $\{X_1, \ldots, X_n\}$ be a random sample of **normal** observable R.V.'s with **unknown mean** μ_X and **unknown variance** σ^2.
  - Let $\overline{X}_n$, s_X^2 be the sample mean and the **unbiased sample variance** respectively.
- Let $\{Y_1, \ldots, Y_m\}$ be a random sample of **normal** observable R.V.'s with **unknown mean** μ_Y and **unknown variance** σ^2.
  - Let $\overline{X}_n$, s_X^2 be the sample mean and the **unbiased sample variance** respectively.
- Here, we assume every X_i and Y_j have the same variance σ^2.

Definition: The two-sample t-statistic of $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ is the R.V.

$$T = \frac{\sqrt{n+m-2}(\overline{X}_n - \overline{Y}_m)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{(n-1)s_X^2 + (m-1)s_Y^2}}.$$

Theorem: If $\mu_X = \mu_Y$, then the **two-sample t-statistic** has the t-distribution with $m+n-2$ degrees of freedom.

Two-sample t-test

Definition: A two-sample t-test is a t-test that uses the two-sample t-statistic (or its absolute value) as the test statistic.

Three most important examples of two-sample t-tests:
Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ be two random samples of **normal** observable R.V.'s, where each X_i has **unknown mean** μ_X, each Y_j has **unknown mean** μ_Y, and all of the X_i's and Y_j's have a **common unknown variance** σ^2. Let $c \in \mathbb{R}$, and let T be the **two-sample t-statistic** of $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$.

- The t-test with null hypothesis $H_0 : \mu_X \leq \mu_Y$, test statistic T, and rejection region $[c, \infty)$ is a two-sample t-test.
- The t-test with null hypothesis $H_0 : \mu_X \geq \mu_Y$, test statistic T, and rejection region $(-\infty, c]$ is a two-sample t-test.
- The t-test with null hypothesis $H_0 : \mu_X = \mu_Y$, test statistic T, and rejection region $[c, \infty)$ is a two-sample t-test.

Note: The first two two-sample t-tests are called one-sided, while the third two-sample t-test is called two-sided.

Significance levels and p-values of two-sample t-tests

Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ be two random samples of **normal** observable R.V.'s, where each X_i has **unknown mean** μ_X, each Y_j has **unknown mean** μ_Y, and all of the X_i's and Y_j's have a **common unknown variance** σ^2. Let $c_0 \in \mathbb{R}$, and let T be the **two-sample t-statistic** of $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$.

Theorem: Suppose that c_0 is the $100(1-\alpha_0)$-percentile of the t-distribution with $n+m-2$ degrees of freedom.

1. Let $\mathcal{H}$ be a t-test with null hypothesis $H_0 : \mu_X \leq \mu_Y$, test statistic T, and rejection region $[c, \infty)$.
   - $\mathcal{H}$ has significance level α_0 if and only if $c \geq c_0$.
   - If the observed value of T is c_0, then the p-value of $\mathcal{H}$ is α_0.
2. Let $\mathcal{H}$ be a t-test with null hypothesis $H_0 : \mu_X \geq \mu_Y$, test statistic T, and rejection region $(-\infty, c]$.
   - $\mathcal{H}$ has significance level α_0 if and only if $c \leq c_0$.
   - If the observed value of T is c_0, then the p-value of $\mathcal{H}$ is α_0.
3. Let $\mathcal{H}$ be the t-test with null hypothesis $H_0 : \mu_X = \mu_Y$, test statistic $|T|$, and rejection region $[c, \infty)$.
   - $\mathcal{H}$ has significance level $2\alpha_0$ if and only if $c \geq c_0$.
   - If the observed value of $|T|$ is c_0, then the p-value of $\mathcal{H}$ is $2\alpha_0$.

χ^2 statistic

Let $\{X_1, \ldots, X_n\}$ be a random sample of observable R.V.'s.

- Each X_i has k possible values, representing k possible types.
- θ_i is the unknown probability that type i is selected.
- $p_1, \ldots, p_k$ are given real numbers, representing our guess for the actual values of $\theta_1, \ldots, \theta_k$.
- After the observed values $X_1 = x_1, \ldots, X_n = x_n$ have been obtained, let N_i be the number of observed values of type i.

Consider a hypothesis test with the following hypotheses:

$$H_0 : \theta_i = p_i \text{ for all } i \in \{1, \ldots, k\} \quad \text{(null hypothesis);}$$

$$H_1 : \theta_i \neq p_i \text{ for at least one } i \quad \text{(alternative hypothesis).}$$

Definition: The χ^2 statistic is a statistic of $\{X_1, \ldots, X_n\}$ given by

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}.$$

Important Theorem: If H_0 is true and the sample size $n \to \infty$, then Q converges in distribution to the χ^2 **distribution** with $k - 1$ degrees of freedom.

χ^2 test of goodness of fit

Definition: A χ^2 test of goodness of fit (or simply, a χ^2 test) is a hypothesis test $\mathcal{H}$ on categorical data that satisfies the following:

- The data has k categories. A random sample of size n is selected from the data. (Typically, the sample size n is large.)
- $\theta_i = \Pr(\text{randomly selected data point is in category } i)$.
- The null hypothesis of $\mathcal{H}$ is $H_0 : (\theta_1, \ldots, \theta_k) = (p_1, \ldots, p_k)$.
- The test statistic is the $\boldsymbol{\chi^2}$ **statistic**, with rejection region $[c, \infty)$.

When to use χ^2 test?

- The χ^2 test is a test for **goodness of fit**.
  - i.e. test how well the data "fits" our null hypothesis.
- **Note:** The alternative hypothesis has no assumption on the distribution of the data.
- **Interpretation:** We use the χ^2 test to see if our "guess" distribution is reasonable. In contrast, in the usual hypothesis test, we are testing if a specific range of values is a reasonable "guess" for the values of some parameters (of some given fixed family of distributions).

Summary

- Statistical model, parameter space, statistic
- Prior and posterior distributions
- Families of conjugate priors
- Estimators and Estimates
- Special distributions (beta, gamma, χ^2, t)
- Useful results on χ^2 distribution and t-distribution
- Confidence intervals and hypothesis testing
- Error, significance level, power, p-value
- Likelihood statistic, likelihood ratio test, likelihood ratio
- t-test, two-sample t-test
- χ^2 test of goodness of fit

Due to the lack of time, we shall review the least squares method and linear regression during cohort class this week.

Reminder: The **Final Exam** will be held on 3rd May (Friday), 9–11am, at the **Indoor Sports Hall 2** (61.106).

- Tested on all materials covered in this course. 1 piece of A4-sized double-sided **handwritten** cheat sheet is allowed.