S&DS 242/542: Theory of Statistics

Lecture 6: z-tests and t-tests for one and two samples

One-sample tests

Given a data sample $X_1, \ldots, X_n$, is their distribution "centered around 0"?

Example: A diagnostic exam is given at the start of S&DS 242, and a second exam is given at the end. Based on differences in student scores, did S&DS 242 improve their knowledge of statistics?

Let $X_1, \ldots, X_n$ be the increases in exam scores for n students. Assume these are IID from a common distribution with mean μ. One way to formulate this as a testing problem is:

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_1 : \mu > 0$$

One may consider more *parametric* formulations that specify a precise form for the distribution of $X_1, \ldots, X_n$, formulations that test for the median instead of the mean, etc.

Two-sample tests

Given two data samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, are their distributions different? Is one distribution "larger than" the other?

Example: A student criticizes the preceding setup. What if the professor made the second exam easier, to show an improvement? To address this criticism, we give the exams to a *control group* of students not taking S&DS 242.

Let $X_1, \ldots, X_n$ be the score increases for the S&DS 242 students, assumed IID from a distribution with mean μ_X. Let $Y_1, \ldots, Y_m$ be those of the control group, assumed IID from a distribution with mean μ_Y. One way to formulate the testing problem is:

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad H_1 : \mu_X > \mu_Y$$

One may again consider more parametric formulations, tests for other differences in distribution beyond the mean, etc.

One-sample z-test and t-test

One-sample z-test

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. We wish to test

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_1 : \mu > 0$$

This is a formulation of the one-sample testing problem, assuming a normal model for the data.

Suppose first that σ^2 is known. It is natural to consider the **z-statistic**

$$Z = \frac{\sqrt{n}}{\sigma} \cdot \bar{X}$$

where $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$ is the sample mean. Large values of Z provide evidence against H_0 in favor of H_1.

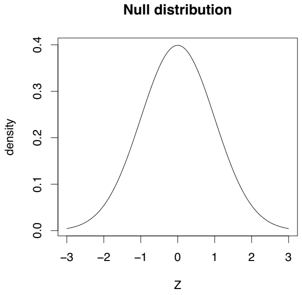
Null distribution of z-statistic

To determine how to perform the test, we need to understand the **null distribution** of Z:

Under H_0: $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$

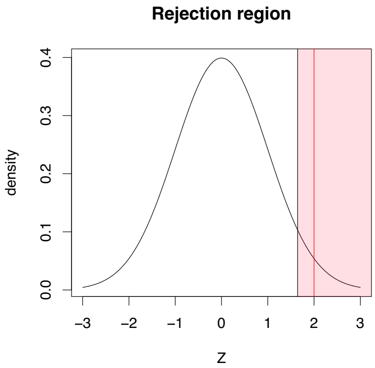
$\Rightarrow \mathbb{E}\bar{X} = 0, \quad \text{Var}\,\bar{X} = \frac{\sigma^2}{n}, \quad \bar{X} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$

$\Rightarrow Z = \frac{\sqrt{n}}{\sigma} \cdot \bar{X} \sim \mathcal{N}(0, 1)$



One-sample z-test

The one-sample z-test at **significance level** α would reject H_0 when $Z > z^{(\alpha)}$, the upper-α point of the distribution $\mathcal{N}(0, 1)$.



One-sample z-test

The **p-value** of this one-sample z-test is the right tail probability of the standard normal distribution at Z,

$$P = 1 - \Phi(Z) = \Phi(-Z)$$

where $\Phi(\cdot)$ is the standard normal CDF.

If we test against $H_1 : \mu < 0$ and reject H_0 for small (i.e. large negative) values of Z, the p-value would be the left tail probability

$$P = \Phi(Z)$$

If we test against $H_1 : \mu \neq 0$ and reject H_0 for large values of $|Z|$, the p-value would be the two-sided tail probability

$$P = (1 - \Phi(|Z|)) + \Phi(-|Z|) = 2\Phi(-|Z|)$$

One-sample t-test

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. We wish to test

$$H_0 : \mu = 0 \qquad \text{vs.} \qquad H_1 : \mu > 0$$

Usually the variance σ^2 is not known and must be estimated from the data. A standard estimator for σ^2 is the **sample variance**

$$S^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n - 1}$$

Substituting S for σ in the z-statistic gives the **(one-sample) t-statistic**

$$T = \frac{\sqrt{n}}{S} \cdot \bar{X}$$

The test of H_0 vs. H_1 based on T is the **one-sample t-test**.

The sample variance

Theorem

If $X_1, \ldots, X_n$ are IID from any distribution with finite variance σ^2, then $S^2 = \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}$ is an unbiased estimator of σ^2.

Proof: Assume $\mathbb{E}X_i = 0$. Expand

$$(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2 = (X_1^2 - 2X_1\bar{X} + \bar{X}^2) + \ldots + (X_n^2 - 2X_n\bar{X} + \bar{X}^2)$$

$$= X_1^2 + X_2^2 + \ldots + X_n^2 - 2\underbrace{(X_1 + X_2 + \ldots + X_n)}_{= n\cdot\bar{X}}\bar{X} + n\cdot\bar{X}^2$$

$$= X_1^2 + \ldots + X_n^2 - n\bar{X}^2$$

$$\mathbb{E}[X_i^2] = \mathrm{Var}[X_i] = \sigma^2, \quad \mathbb{E}[\bar{X}^2] = \mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

$$\Rightarrow \mathbb{E}[S^2] = \frac{1}{n-1}\mathbb{E}[X_1^2 + \ldots + X_n^2 - n\bar{X}^2]$$

$$= \frac{1}{n-1}\left(\sigma^2 + \ldots + \sigma^2 - n\cdot\frac{\sigma^2}{n}\right) = \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2.$$

Sampling distribution of the sample variance

Theorem

If $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, then S^2 is independent of $\bar{X}$, with distribution

$$S^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$$

It suffices to prove this for $\mu = 0$ and $\sigma^2 = 1$. (Why?)

Let $Z_1, \ldots, Z_n \overset{IID}{\sim} \mathcal{N}(0,1)$, $\bar{Z}$ and S_Z^2 be their sample mean/variance.

Suppose $\bar{Z}$ is independent of S_Z^2 and $S_Z^2 \sim \frac{1}{n-1} \chi^2_{n-1}$.

Represent $X_i = \mu + \sigma Z_i \Rightarrow \bar{X} = \mu + \sigma \cdot \bar{Z}$, $S^2 = \sigma^2 \cdot S_Z^2$

$\Rightarrow S^2$ independent of $\bar{X}$, and $S^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$

Sampling distribution of the sample variance

Proof #1: $\bar{X}$ is independent of S^2, Assume $\mu=0$, $\sigma^2=1$.

- Consider $(\bar{X}, X_1-\bar{X}, X_2-\bar{X}, \ldots, X_n-\bar{X})$. This is linear transformation of $(X_1,\ldots,X_n)$, so it's MVN.
- $\mathrm{Cov}[\bar{X}, X_i-\bar{X}] = \mathrm{Cov}[\bar{X}, X_i] - \mathrm{Cov}[\bar{X},\bar{X}]$

$$\mathrm{Cov}[\bar{X}, X_i] = \mathrm{Cov}\left[\frac{1}{n}(X_1+\ldots+X_n), X_i\right]$$

$$= \frac{1}{n}\sum_{j=1}^{n} \underbrace{\mathrm{Cov}[X_j, X_i]}_{=0 \text{ unless } j=i} = \frac{1}{n}\mathrm{Var}[X_i] = \frac{1}{n}$$

$$\mathrm{Cov}[\bar{X},\bar{X}] = \mathrm{Var}[\bar{X}] = \frac{1}{n} \Rightarrow \mathrm{Cov}[\bar{X}, X_i-\bar{X}] = 0.$$

$\Rightarrow \bar{X}$ is uncorrelated with each of $X_1-\bar{X}, \ldots, X_n-\bar{X}$.

- Implies $\bar{X}$ independent of $(X_1-\bar{X}, \ldots, X_n-\bar{X})$ b/c MVN. $\Rightarrow \bar{X}$ independent of S^2.

Sampling distribution of the sample variance

Proof #2: $\bar{X}$ is independent of S^2. Assume $\mu=0$, $\sigma^2=1$.

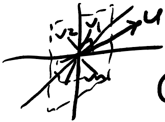
- Let $X=(X_1,\dots,X_n)\in\mathbb{R}^n$, standard MVN.

$$\bar{X} = \underbrace{\left(\tfrac{1}{n},\dots,\tfrac{1}{n}\right)}_{=u}{}^T X, \qquad X_1-\bar{X} = \underbrace{\left(1-\tfrac{1}{n}, -\tfrac{1}{n},\dots,-\tfrac{1}{n}\right)}_{=v_1}{}^T X$$

- $$u^T v_1 = \tfrac{1}{n}\left(1-\tfrac{1}{n}\right) + \underbrace{\tfrac{1}{n}\left(-\tfrac{1}{n}\right)+\dots+\tfrac{1}{n}\left(-\tfrac{1}{n}\right)}_{n-1 \text{ times}} = \tfrac{1}{n} - n\cdot\tfrac{1}{n^2} = 0$$

$\Rightarrow$ u is perpendicular to v_1. Length of $u = \frac{1}{\sqrt{n}}$.

Define $X_i-\bar{X} = v_i^T X$. Then same for u and v_i.



By rotational symmetry of MVN, $\bar{X}$ and $(X_1-\bar{X},\dots,X_n-\bar{X})$ have same distribution as $\frac{1}{\sqrt{n}}X_1$ and some combinations of $X_2,\dots,X_n$

Sampling distribution of the sample variance

Proof: $S^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$. Assume $\mu = 0$, $\sigma^2 = 1$.

- By previous calculation:

  $$S^2 = \frac{1}{n-1}\left[X_1^2 + \dots + X_n^2 - n\bar{X}^2\right]$$

- Let $U = (n-1)S^2$, $W = X_1^2 + \dots + X_n^2$, $V = n \cdot \bar{X}^2$.

  $\Rightarrow U = W - V \Rightarrow W = U + V$

- S^2, $\bar{X}$ are independent, so also U, V independent.

- $W \sim \chi^2_n$, $\sqrt{n}\bar{X} \sim N(0,1) \Rightarrow V = (\sqrt{n}\bar{X})^2 \sim \chi^2_1$

  $\Rightarrow$ This implies $U \sim \chi^2_{n-1}$.

  $\Rightarrow S^2 = \frac{1}{n-1} U \sim \frac{1}{n-1} \chi^2_{n-1}$.

The Student-t distribution

Definition

The distribution of $T = Z/\sqrt{U}$ when Z, U are independent random variables with $Z \sim \mathcal{N}(0,1)$ and $U \sim \frac{1}{n} \cdot \chi^2_n$ is called the **Student-t distribution** with n degrees of freedom. We write

$$T \sim t_n$$

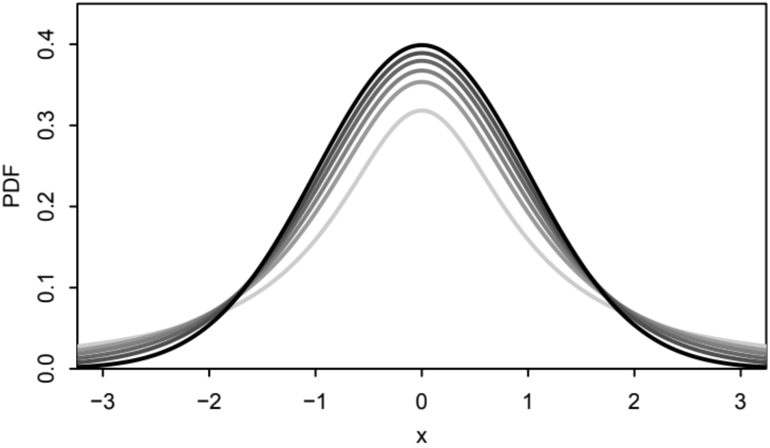
The one-sample t-statistic is

$$T = \frac{\sqrt{n}}{S} \cdot \bar{X} = \underbrace{\frac{\sqrt{n}}{\sigma} \cdot \bar{X}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{S^2}{\sigma^2}}_{=U}}$$

Under $H_0 : X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, Z and U are independent with $Z \sim \mathcal{N}(0,1)$ and $U \sim \frac{1}{n-1} \cdot \chi^2_{n-1}$. So the null distribution of T is

$$T \sim t_{n-1}$$

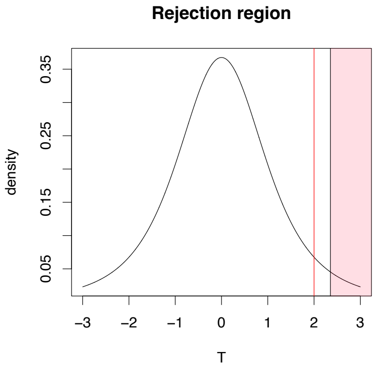
The Student-t distribution



PDF of the t-distribution t_n with degrees-of-freedom $n = 1, 2, 3, 5, 10$ (light to dark)

One sample t-test

The one-sample t-test at significance level α would reject H_0 when $T > t_{n-1}^{(\alpha)}$, the upper-α point of the distribution t_{n-1}.



Composite hypotheses and pivotal statistics

Note that if σ^2 is unknown in the null hypothesis

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$$

then H_0 describes a family of distributions instead of a single distribution. Such a hypothesis is **composite** rather than **simple**.

For a composite null hypothesis H_0, a level-α test must satisfy

$$\mathbb{P}_{H_0}[\text{Type I error}] = \mathbb{P}_{H_0}[\text{reject } H_0] \leq \alpha$$

under *every* distribution specified by H_0.

Luckily, T has the same distribution for every $\sigma^2 > 0$, i.e. under every distribution specified by H_0, so we can define the same rejection threshold. Such a statistic is called **pivotal** under H_0.