Review of Basic Probability and Statistics

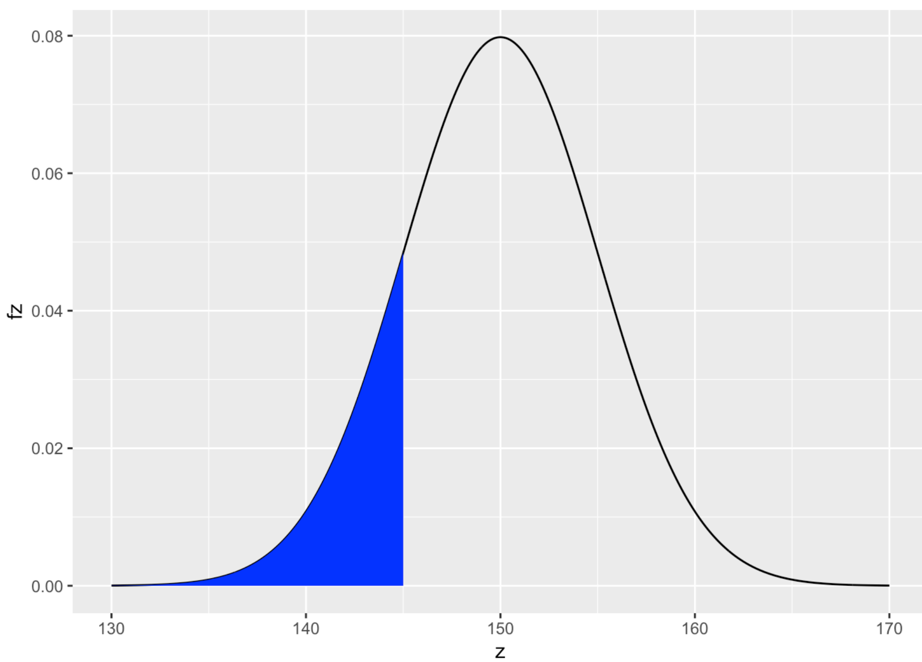
Hao Zhang

Review of Probability

Random Variables

A random variable such as your weight takes its values by chance. It is described by a probability distribution.

A continuous random variable is one whose values may fill an interval. Its probability distribution is described by its probability density function (pdf).The probability that the random variable falls into an interval is the area under the curve.



Random variables are usually denoted by capital letters X, Y, Z.

Independence

Two random variables X and Y are independent if

condition　　　uncondition

$$P(Y \le b | X \le a) = P(Y \le b)$$

“condition probability of Y given X”

- If X and Y are independent, the distribution of Y doesn't change no matter what value X holds.
- Conditioning on $X \le a$ does nothing to Y's distribution.

for any values a and b.

Mean and Variance

central tendency

The mean of Y, denoted by $E(Y)$ or μ, is the center of the probability distribution. The variance of Y is defined by

$$Var(Y) = E[(Y - \mu)^2].$$

spread out of Y

Hence

Alternative formula

$$Var(Y) = E(Y^2) - \mu^2.$$

Properties

1. $E(a_1 Y_1 + a_2 Y_2) = a_1 E(Y_1) + a_2 E(Y_2)$ for any constants a_1 and a_2. Linearity of Expectation
2. $\mathrm{Var}(a_1 Y_1) = a_1^2 \mathrm{Var}(Y_1)$. Variance of Scaled Variable
3. $\mathrm{Var}(a_1 Y_1 + a_2 Y_2) = a_1^2 \mathrm{Var}(Y_1) + a_2^2 \mathrm{Var}(Y_2) + 2 a_1 a_2 \mathrm{Cov}(Y_1, Y_2)$. Variance of Linear Combination → when variables are not independent, must include covariance.
4. If Y_1 and Y_2 are independent or uncorrelated, then $\mathrm{Var}(a_1 Y_1 + a_2 Y_2) = a_1^2 \mathrm{Var}(Y_1) + a_2^2 \mathrm{Var}(Y_2)$.

no covariance. (=0)

The above properties generalizes to the sum of n variables. For example

Mean (Expected Value):

$$E(\sum_{i=1}^{n} a_i Y_i) = \sum_{i=1}^{n} a_i E(Y_i).$$

The mean of weighted sum = sum of weights × means of each variable.

Quantity	Formula
Mean	$E\left(\sum a_i Y_i\right) = \sum a_i E(Y_i)$
Variance (independent)	$\mathrm{Var}\left(\sum a_i Y_i\right) = \sum a_i^2 \mathrm{Var}(Y_i)$

- Y_i = random variable (like a test score, price, etc.)
- a_i = constant weight (like a percentage or multiplier)
- n = total number of variables

Special distributions

Normal distributions

likelihood of observing value x.

Probability Density Function (pdf)

The $N(\mu, \sigma^2)$ has a pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

where μ is the mean and σ^2 the variance. The normal distribution is denoted by $N(\mu, \sigma^2)$. $N(0, 1)$ is called the standard norm.

- If Y is $N(\mu, \sigma^2)$, $a + bY$ is $N(a + b\mu, b^2\sigma^2)$. Linear Transformation
- If Y is $N(\mu, \sigma^2)$, $(Y - \mu)/\sigma \sim N(0, 1)$. This is called the standardization. Standard normal $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- If Y_1 is $N(\mu_1, \sigma_1^2)$ and Y_2 is $N(\mu_2, \sigma_2^2)$ and the two variables are independent, then $b_1Y_1 + b_2Y_2$ is $N(b_1\mu_1 + b_2\mu_2, b_1^2\sigma_1^2 + b_2^2\sigma_2^2)$.

Sum of Independent Normals

The sum of independent normal random variables is a normal random variable.

χ^2 distributions

$Z^2 = \chi^2$

For practical purpose, we use the following fact as the definition and also a property of χ^2 distribution:

If $Z_1, \ldots, Z_n$ are independent standard normal random variables, then $\sum_{i=1}^{n} Z_i^2$ has the χ^2 distribution with n degrees of freedom. iid

If Y_1 has χ^2_m distribution and Y_2 has the χ^2_n distribution and the two variables are independent, then $Y_1 + Y_2$ has the χ^2_{m+n} distribution.

Cochran's theorem (regression/ANOVA)

If $Z_1, \ldots, Z_k$ are independent identically distributed (i.i.d.), standard normal random variables, then $Q = \sum_{i=1}^{k} \left(Z_i - \bar{Z}\right)^2 \sim \chi^2_{k-1}$ where

dof

$$\bar{Z} = \frac{1}{k}\sum_{i=1}^{k} Z_i$$

In addition, Q and $\bar{Z}$ are independent.

If $X_1, \ldots, X_n$ are i.i.d. $N\left(\mu, \sigma^2\right)$ random variables, then

$$V = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sim \chi^2_{n-1}$$

Variance

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and V and $\bar{X}$ are independent.

t-distribution

Accounts for extra variability when population variance is unknown and estimated from data.

① Small samples
② Wider (heavier tails)
③ Inference on population mean when variance is unknown.

Estimating mean when σ² unknown

The t-distribution with ν degrees of freedom can be defined as the distribution of the random variable T with

$$T = \frac{Z}{\sqrt{V/\nu}}$$

where

- Z is a standard normal random variable;
- V has a chi-squared distribution (χ^2-distribution) with ν degrees of freedom;
- Z and V are independent.

The t-distribution is essential for the inferences about the mean of a normal distribution.

F-distribution

① ANOVA
② Model Comparison
③ testing if pop. have same variance.

Comparing variances or model fits

The F-distribution with d_1 and d_2 degrees of freedom is the distribution of ratio

$$X = \frac{S_1/d_1}{S_2/d_2}$$

where S_1 and S_2 are independent random variables with chi-square distributions with respective degrees of freedom d_1 and d_2.

The F-distribution is essential for comparing the means of multiple normal distributions.

Review of Statistics

Sampling distributions

Let $Y_1, \dots, Y_n$ be a random sample from $N(\mu, \sigma^2)$, and let

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

be the sample mean and sample variance.

The distributions of sample statistics such as the sample mean and sample variance are called sampling distributions.

- $\bar{Y}$ and $(Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})$ are independent.
- $\bar{Y}$ and S^2 are independent.
- $\bar{Y} \sim N(\mu, \sigma^2/n)$.
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Infer pop. var.
- t-statistic $\boxed{\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}}$. "$\sigma$" is unknown: Estimated by "S"

These properties are extremely important for statistical inferences.

Hypothesis Testing

For one sample mean

$$H_0 : \mu = \mu_0,\ H_1 : \mu \neq \mu_0.$$

Data: $y_1, \dots, y_n$.

Test statistic:

$$\boxed{t = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}}.$$

where $\bar{Y}$ and S are the sample mean and sample standard derivation, respectively.

Intuition: Reject H_0 is t is too big or too small.

P-value:

$$p = 2P(T > |t|)$$

where T has a t-distribution with $n-1$ degrees of freedom, and t is the value of the test statistic.

Two-sample means

Observed two samples from two normal distributions and test if the means are equal

- Sample one from $N(\mu_1, \sigma^2)$:

$$Y_{11}, Y_{12}, \dots, Y_{1,n_1}.$$

  - Sample statistics: $\bar{Y}_1$ and S_1^2.
- Sample two from $N(\mu_2, \sigma^2)$:

$$Y_{21}, Y_{22}, \dots, Y_{2,n_2}.$$

  - Sample statistics: $\bar{Y}_2$ and S_2^2.
- Note the variances are assumed to be equal but unknown.
- Hypothesis:

$$H_0 : \mu_1 = \mu_2,\ H_1 : \mu_1 \neq \mu_2.$$

- Test statistic

$$\boxed{t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

For: Equal Variances ($\sigma_1^2 = \sigma_2^2$)
(use Sp)

where

✅ 2. Unequal variances (Welch's t-test)

Use when you **cannot assume equal variances**.

Test statistic:

$$\boxed{t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}}$$

Degrees of freedom (Welch–Satterthwaite approximation):

$$\boxed{df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}}$$

pooled variance $\boxed{S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$

Under H_0 it has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. To see this, observe

- Distribution of the difference:

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)\right).$$

- Distribution of the sample variances:

$$\begin{aligned}(n_1-1)S_1^2/\sigma^2 &\sim \chi^2_{n_1-1}\\ (n_2-1)S_2^2/\sigma^2 &\sim \chi^2_{n_2-1}\\ \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} &\sim \chi^2_{n_1+n_2-2}\end{aligned}$$

- P-value

$$p = 2P(T > |t|)$$

where $T \sim t_{n_1+n_2-2}$.

Revisit of the two-sample test

Given two samples from two normal populations, consider testing

$$H_0: \mu_1 = \mu_2,\ H_1: \mu_1 \neq \mu_2.$$

We can rewrite the t-test in a form that can be generalized to testing for more than two normal means.

Write

$$\begin{aligned}\bar{y}_{..} &= \tfrac{1}{n_1+n_2}\left(\textstyle\sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j}\right) = \tfrac{1}{n_1+n_2}(n_1\bar{y}_1 + n_2\bar{y}_2)\\ \bar{y}_1 - \bar{y}_{..} &= \tfrac{n_2}{n_1+n_2}(\bar{y}_1 - \bar{y}_2)\\ \bar{y}_2 - \bar{y}_{..} &= \tfrac{n_1}{n_1+n_2}(\bar{y}_2 - \bar{y}_1)\\ n_1\left(\bar{y}_1 - \bar{y}_{..}\right)^2 + n_2\left(\bar{y}_2 - \bar{y}_{..}\right)^2 &= \tfrac{n_1 n_2}{n_1+n_2}(\bar{y}_1 - \bar{y}_2)^2\end{aligned}$$

Between-Group (Numerator)

Then for the two-sample t-test, we have

$$\boxed{F = t^2 = \frac{n_1\left(\bar{y}_1 - \bar{y}_{..}\right)^2 + n_2\left(\bar{y}_2 - \bar{y}_{..}\right)^2}{\frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2}{n_1+n_2-2}}}$$

$df_1 = 1$
$df_2 = n_1 + n_2 - 2$

This test statistic has an F-distribution with 1 and $(n_1 + n_2 - 2)$ degrees of freedom.

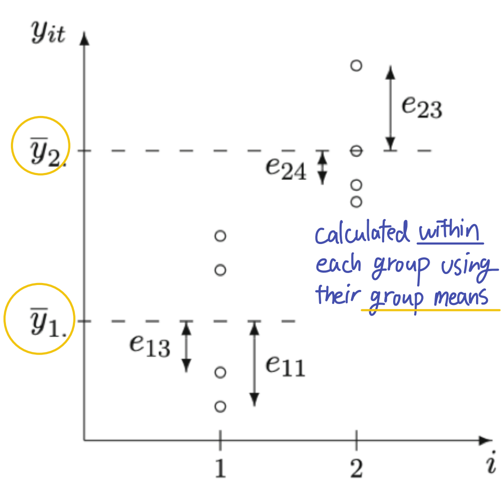
Note

- The numerator represents between-group variation.
- The denominator represents within-group (residual) variation. pooled variance
- The numerator and denominator are independent (why?)
- The following graphs further illustrates the terms.

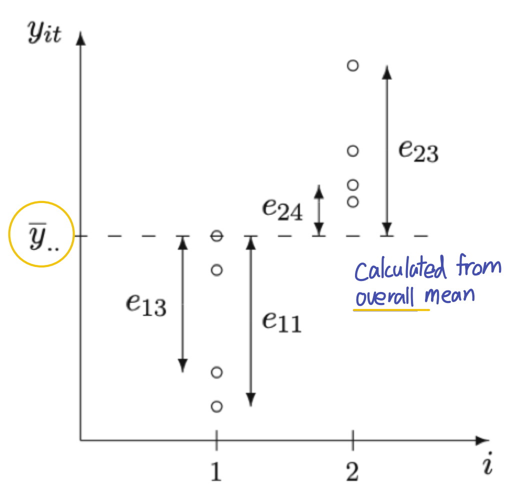
🎯 **Why Independent?**

Because:

- The **numerator** depends only on the sample means $\bar{y}_1$ and $\bar{y}_2$.
- The **denominator** depends on the spread (variability) of individual data **around their own group mean**, not on the means themselves.



Residuals; full model



Residuals; reduced model

✏️ **Special case: Two-sample t-test as an F-test**

When comparing **two group means** (like in your slide):

- The **full model** assumes the two groups can have different means.
- The **reduced model** assumes the groups share the same mean.

Then the F-statistic becomes:

$$F = \frac{\text{Between-group variation}}{\text{Within-group variation}} = t^2$$

✅ **F-test formula:**

$$\boxed{F = \frac{(RSS_{\text{reduced}} - RSS_{\text{full}})/(df_{\text{reduced}} - df_{\text{full}})}{RSS_{\text{full}}/df_{\text{full}}}}$$

📌 **What do these terms mean?**

- RSS = **Residual Sum of Squares**
- df = **Degrees of Freedom**
- **Reduced model:** fewer predictors (simpler)
- **Full model:** more predictors (complex)

- If the **group means** differ substantially from the grand mean, you get **large between-group variation** → stronger evidence to reject H_0.