Contents

Day 14

Continuous Random Variables

Can take any real number ($\mathbb{R}$) value within any given interval.

We cannot use a probability mass function so we will instead use a probability density function (PDF) denoted as $f(x)$

Properties

- The probability of being in an interval (a, b] is:

$$\int_a^b f(x)dx = \int_{-\infty}^b f(x) - \int_{-\infty}^a f(x)dx$$

  - This is considered the area under the curve between a and b
- $P(X = x) = 0 \ \forall x$
  - $P(X \leq x) = P(X < x)$
  - $P(X \geq x) = P(X > x)$

$f(x)$ is displayed graphically as a density curve

Properties of $f(x)$

- $\forall \ x \in \mathbb{R}, \ f(x) \geq 0$
  - Density never goes below x-axis
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Mean of continous random variable: $\mu_x = \int_{-\infty}^{\infty} x \times f(x)dx$

Variance of continous random variable is $\sigma^2 = \int_{-\infty}^{\infty} (X = \mu_x)^2 f(x)dx$

Uniform Random Variable

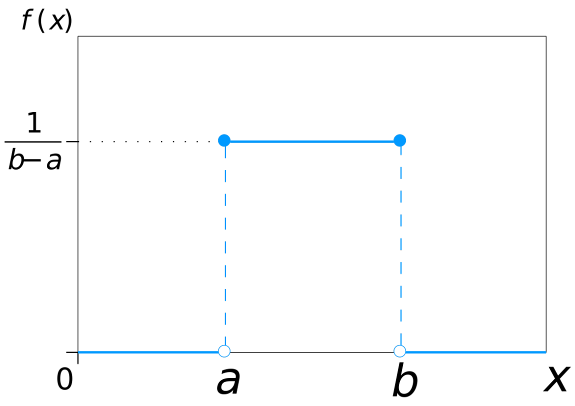

Figure 1: Graphical Representation

$X \sim U(a, b)$

Example : Standard Uniform Random Variable

$X \sim U(0, 1)$

Find

- $\mathrm{P}(\mathrm{X} \geq 0.3)$
- $\mathrm{P}(\mathrm{X} = 0.3)$
- $\mathrm{P}(0.3 < \mathrm{X} \leq 1.3)$
- $\mathrm{P}(0.2 \leq \mathrm{X} \leq$ **or** $0.7 \leq \mathrm{X} \leq 0.9)$
- P(X is not in the interval (0.4, 0.7))

Answers

- $\sqcap = (0.7) \times (1) = 0.7$
- $\sqcap = 0$
  - The probability of being exactly on a point in the infinite sum will **always** be 0.
- $\sqcap = (0.7) \times (1) = 0.7$
  - Do not keep shading when there is no density curve, meaning it is a hard stop at $X = 1$
- $\sqcap = ((0.25 - 0.2) \times \frac{1}{0.25-0.2}) + ((0.91 - 0.7) \times \frac{1}{0.91-0.7}) = 0.25$
- $\sqcap = 0.4 + 0.3 = 0.7$

Normal Random Variable

$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Density curve is also a "bell curve"

Empirical (68-95-99.7) Rule

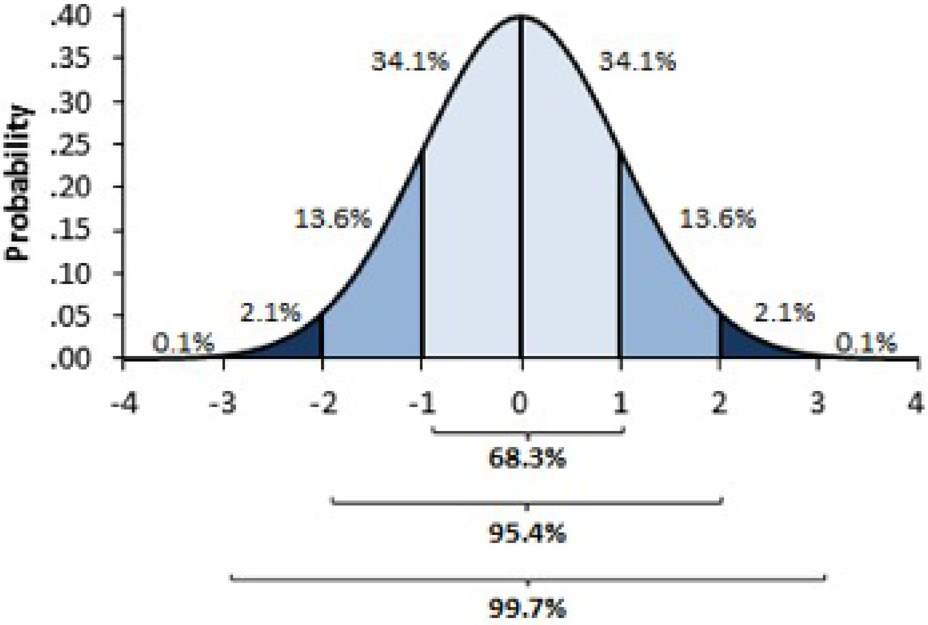

Figure 2: Bell Curve

$X \sim N(\mu\,, \sigma)$

Standardization

It may be useful to standardize distributions to compare 2 variables with same density curve shape but different scales. For normal distributions $X \sim N(\mu, \sigma)$, we convert to Z-Scores $Z \sim N(0, 1)$

$$Z = \frac{x - \mu}{\sigma} = \frac{value\, -\, mean\, of\, distribution}{standard\, deviation}$$

$$P(Z \leq z) = P(X \leq x)$$

"Cumulative proportion"/"Cumulative probability"

Contents

Day 15

Z-Score Example

Two tests of "English" ability

- NAEP Reading Test
- SAT verbal

Suppose a student score scored 320 on NAEP & 650 SAT

Which test did he do better on?

$NAEP \sim N(288, 38)$

$SAT \sim N(500, 120)$

Convert to Z-Scores

NAEP:

$$Z = \frac{value\ -\ mean}{standard\, deviation} = \frac{320 - 288}{38} = 0.842$$

Student scored 0.842 standard deviation above average

SAT:

$$Z = \frac{value\ -\ mean}{standard\, deviation} = \frac{650 - 500}{120} = 1.25$$

Student scored 1.25 standard deviation above average

R-Code

```
pnorm(320, mean = 288, sd = 38)
[1] 0.8001355
```

Cumulative proportion of 0.800 (80%) which means 80th percentile.

```
pnorm(650, mean = 500, sd = 120)
[1] 0.8943502
```

Cumulative proportion of 0.8943502 which means 89th percentile.

Water bottle example

Questions

- Why does it continue to overfill
  - How much does it actually pour → average
- Why does Dr. Wynne have such terrible reaction speed?
  - Reaction speed → average
- Does the water fill at the same rate
  - Average rate for one pour
  - → average over several attempts

Expected value = μ = expected amount filled

$\bar{X}$ = average amount filled in a sample of pours "sample mean".

<u>Variability:</u> how variable are the individual values. (range)

- σ = Standard Deviation
- σ^2 = Variance
- S = Sample Standard Deviation
- S^2 = Sample Variance

<u>Bias:</u> <u>Center:</u> - on average, are we where we expected to be? (mean, median, mode)

Shape

Shape: where "average" is compared to "most likely"

- How "consistent" the values are given variability

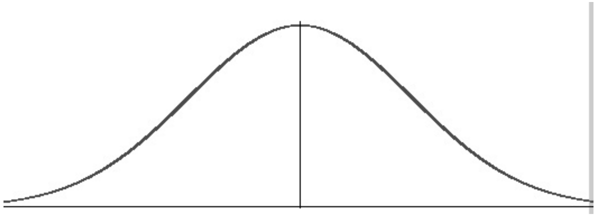

Figure 1: Unimodal Distribution

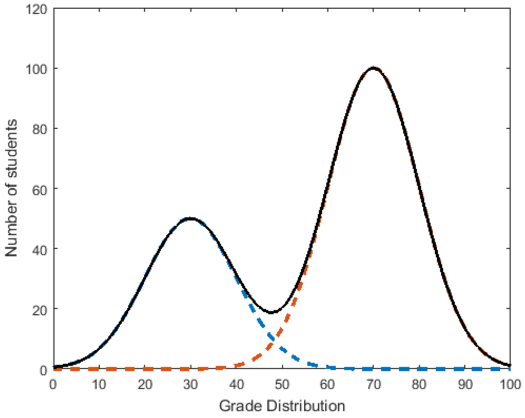

Figure 2: Bimodal Distribution

The median is resistant, the mean is subject to more change.

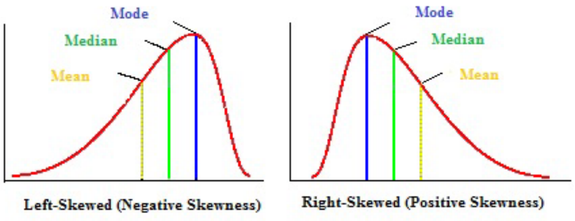
Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Figure 3: Left and Right Skewed Graphs

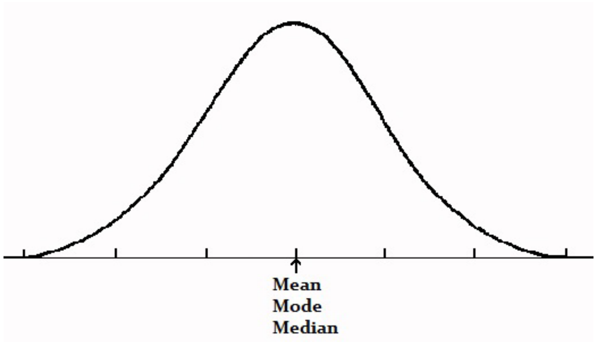

Figure 4: Symmetric Graph

Approximating a Density Curve: Histogram

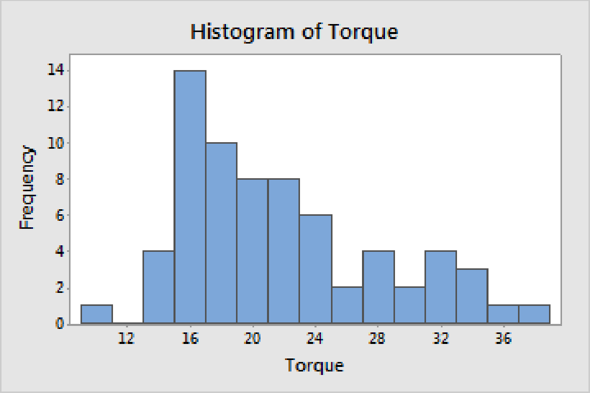

Figure 5: Histogram

- "bins": intervals on the x-axis
- Choice of bins is very important
  - Endpoints of bins
  - Center & Width
- Reimann Integral of an unknown density curve

Outliers

Points that doesn't fit with everything else

Attempting to determine outliers

- Plot your data & look for points that don't belong
- ↑ best way
- Investigate why they're different

Box-Plots

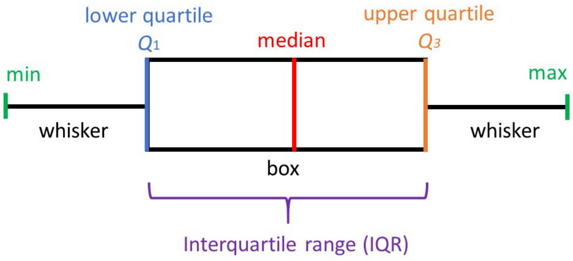

Figure 6: Box Plot

Rule of Thumb

- Step 1: Get five number summary (min, Q_1, medium(Q_2), Q_3, max)
- Step 2: Compute IQR = range middle 50% of data
  - IQR = Q_3 - Q_1
- Step 3: Compute "fences"
  - Lower fence: $Q_1 - K \times IQR$
  - Upper fence: $Q_3 + K \times IQR$

Anything outside the fences is an outlier.

By convention, $k = 1.5$

Example : Senator Ages

Five number summary:

- Min = 39
- Q_1 = 55.5
- Median = 63
- Q_3 = 69
- Max = 85

$IQR = 69 - 55.5 = 13.5$

Lower fence: $55.5 - (1.5)(13.5) = 35.25$ Upper fence: $69 + (1.5)(13.5) = 89.25$

In this data set we have no outliers because our data falls between the fences.

Numerical Variable Connection to Random Variables

Recall for random variable X

$E(A + Bx) = a + b \times E(x)$

$Var(A + Bx) = b^2 \times Var(x)$

$SD(A + Bx) = |b| \times sd(x)$

Recall for random variable X and Y

$E(Ax + By) = aE(x) + bE(y)$

$Var(Ax + By) = A^2 \times var(x) + B^2 var(y)$

$SD(Ax + By) = \sqrt{A^2 \times var(x) + B^2 \times var(y)}$

All of these rules hold for numerical variables too

Contents

Day 16

Error and Variability

Rounding variability: error due to precision of our machine/scale/etc.

1) Never measure exact, only to some tolerance
   - → typically rounding error $\sim (E, -E)$

Example: weight is 150 pounds - reality → $150 \pm U(-0.5, 0, 5)$

2) When making repeated measurements of something, there will be some natural variability, due to many small sources of error. Usually (as long as errors are on the same scale), we can make measurement error of $\sim N(0, \sigma)$

3) Sampling error: error due to only having a sample from the population. Estimate a population mean μ based on a sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_{\mathrm{i}}$$

What is the distribution of $\bar{X}$ over all possible samples = sampling distribution of $\bar{x}$ which is the sampling mean

Consider simple random sample from a given population, the values $\mathrm{x}_1 \rightarrow \mathrm{x}_\mathrm{n}$ values of some numerical variables. We assume the x_i's are independent and identically distributed random variables.

With theoretical/population mean μ and standard deviation σ, then $\bar{X} = \frac{1}{n} \rightarrow (X_1 + X_2 + ... + X_\mathrm{n})$ is considered a linear combination.

$$E(\bar{X}) = E(\frac{1}{n} \times (X_1 + X_2 + ... + X_\mathrm{n})) = \frac{1}{n}(E(x_1) + E(x_2) + ... + x_\mathrm{n})$$

$$E(\bar{X}) = n\mu \times \frac{1}{n} = \mu$$

Important Facts

- The mean of the sampling distribution of $\bar{X}$ is equal to the population mean μ

$$(Var\bar{X}) = var(\frac{1}{n} \times (x_1 + x_2 + ... + x_\mathrm{n}))$$

$$= (\frac{1}{2})^2 \times [var(x_1 + x_2 + ... + x_\mathrm{n}]$$

$$= (\frac{1}{n})^2 \times [var(x_1) + var(_2) + ... + var(x_\mathrm{n})]$$

- The variance of sampling distribution of $\bar{X}$ is smaller than the population variance by a factor of n. The standard deviation is smaller by a factor of $\sqrt{n}$. Consider a normally distributed population. Theorem: any linear combination of normal random variables is also normally distributed.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Central Limit Theorem

For a simple random sample (ASRS) of size `n` from a population with finite mean μ and finite standard deviation σ:

When `n` is "large enough", $\bar{x}$ is approximately $\sim N(\mu, \frac{\sigma}{\sqrt{n}})$

What does "large enough" depend on?

- How good the approximation needs to be (robust procedures-approximation just needs to be OK)
- Shape of population distribution
  - Higher skew requires larger `n`
  - Outliers in sample suggest larger `n` is needed

Consider a normally distributed population.

$\bar{X}$ is a linear combination of random variables $X_{\text{i}} \sim N(\mu, \sigma)$

Example

You take a sample size of 64 from a population normally distributed with mean of 82 and standard deviation of 24.

a) Find the sampling distribution of the sample mean $\bar{X}$
b) Middle 95% of values of x are expected to be in what interval.
c) Middle 95% of sample means $\bar{X}$ are expected to be in what interval?

Answers

a) $\bar{X} \sim N(82, \frac{24}{\sqrt{64}}) \sim N(82, 3)$

b) (34, 130)

c) $E[\bar{x}] = \mu = 82$, $SD[\bar{x}] = \frac{\sigma}{\sqrt{n}} = \frac{24}{\sqrt{64}} = 3$

$\mu + 2SD[\bar{x}] = 82 + 6 = 88$

$\mu - 2SD[\bar{x}] = 82 - 6 = 76$

The range between 76 and 88

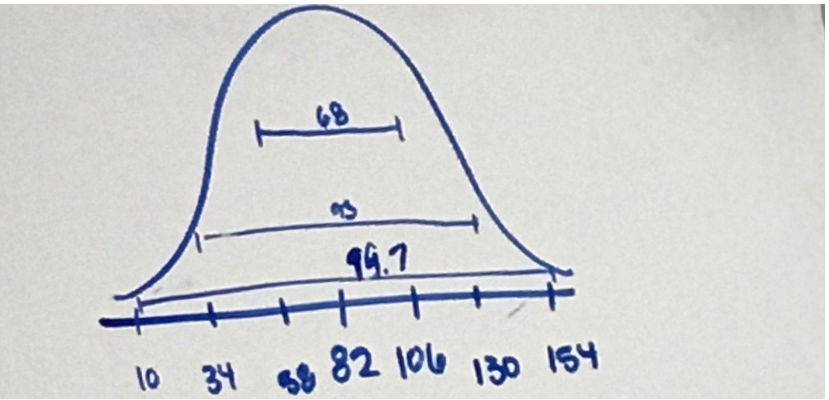

Figure 1: Curve

Contents

Day 17

This material is not on the exam but will tie into different concepts that will be present

Review

Neyman-Pearson Hypothesis Testing

Step 1: Define a parameter and its value under H_0 & H_1

Step 2: Define a test statistic and its sampling distributions under the two hypothesis.

Step 3: Define our alpha level (α) which is $P(Type\, 1\, Error)$. This translates to P(accept H_1 | H_0 is true). Use it to compute a critical region.

Step 4: Collect sample data and compute observed value of test statistic.

Step 5: If value in critical region, then accept H_1. If value not in critical region, then accept H_0.

What changes now?

We are using stand in variables for instead of using categories. This is still using the NPHT framework but uses numerical data.

Step 1: Parameter is μ = population mean. Now we are talking about numerical data now, not categorical data.

- $H_0 = \mu = \mu_0$
- $H_1 = \mu = \mu_1$
- This value of μ is cc

Step 2: Statistic $\bar{X}$ = sample mean. Assuming Central Limit Theorem holds, we will have an approximately normal sampling distribution and will be centered at μ_0 under the null and μ_1 under the alternative hypothesis.

- H_0, $\bar{X} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}})$
- H_1, $\bar{X} \sim N(\mu_1, \frac{\sigma}{\sqrt{n}})$

Sonnet Example

For an old author, sonnets are known to contain an average of 8.9 “new” words, with a standard deviation of 2.5 words. New meaning that they are unique to each of the sonnet and are not used in the others.

A new set of 6 sonnets has been discovered and authorship has been disputed. The sonnets contain an average of 10.2 “new” words. Suppose we believe a different author would use on average 10.9 “new” words.

Using $\alpha = 0.05$, what should we conclude about the authorship of the new sonnets?

Response

Parameter: $\mu =$ population mean which in this context will be the number of “new” words

- Under H_0: $\mu = 8.9$
  - Step 2: $\bar{X} \sim N(8.9, \frac{2.5}{\sqrt{6}})$
- Under H_1: $\mu = 10.9$
  - Step 2: $\bar{X} \sim N(10.9, \frac{2.5}{\sqrt{6}})$
- Here we are stating that it is worse to give credit to someone who didn’t write the sonnets.
- We then accept the null hypothesis because the value is not in the critical region of 10.58. This means the original author wrote the sonnets.
- The way we set up the problem will affect the conclusion we come to.

R-Code

```
alpha <- 0.05
mu <- 8.9
sigma_from_sample <- 2.5/sqrt(6)

qnorm(alpha, mu, sigma_from_sample, lower.tail = FALSE)
```

Neyman-Pearson Power Analysis

Step 1: Define a parameter and its value under H_0 & H_1

Step 2: Define a test statistic and its sampling distributions under the two hypothesis.

Step 3: Define our alpha level (α) which is $P(Type\,1\,Error)$. This translates to P(accept H_1 | H_0 is true). Use it to compute a critical region.

Step 4: Compute our power = P(test statistic in critical region | H_1 is true)

Step 5: [OPTIONAL] Compare power to 80% and/or compute β to $\alpha < \beta \leq 0.2$

Sonnet Example with Power Analysis

For an old author, sonnets are known to contain an average of 8.9 "new" words, with a standard deviation of 2.5 words. New meaning that they are unique to each of the sonnet and are not used in the others.

A new set of 6 sonnets has been discovered and authorship has been disputed. The sonnets contain an average of 10.2 "new" words. Suppose we believe a different author would use on average 10.9 "new" words.

Using $\alpha = 0.05$, what is the power of our test to detect the "new" author?

You are trying to compute a probability here so this R code will suffice

R-Code

```
alpha <- 0.05
old_mu <- 8.9
mu <- 10.9
sigma_from_sample <- 2.5/sqrt(6)

critical_value <- qnorm(alpha, old_mu, sigma_from_sample, lower.tail = FALSE)

pnorm(critical_value, mu, sigma_from_sample, lower.tail = FALSE)
[1] 0.6235198
```

NHST Framework

Step 1: Define a parameter and its value under H_0

Step 2: Define an interval representing an inequality (under H_0, parameter in that interval)

Step 3: Define a test statistic and its sampling distribution under H_0

Step 4: Collect data and compute the observed value of the test statistic.

Step 5: Compute the P-Value which is P(observe a test statistic as or more favorable to H_1 | H_0 is true)

- P(test statistic $\geq$ observed value) OR
- P(test statistic $\leq$ observed value) OR
- P(test statistic "further" from parameter value compared to observed value)

Find either of the three listed above ↑ and double the smaller one.

Step 6: We define our significance level.

If

- P-Value $\leq$ significance level $\implies$ reject H_0 & accept H_1
- P-Value $>$ significance level $\implies$ fail to reject H_0

Sonnet Example with NHST

For an old author, sonnets are known to contain an average of 8.9 "new" words, with a standard deviation of 2.5 words. New meaning that they are unique to each of the sonnet and are not used in the others.

A new set of 6 sonnets has been discovered and authorship has been disputed. The sonnets contain an average of 10.2 "new" words. Suppose we believe a different author would use on average 10.9 "new" words.

Theory suggest a different author would use more "new" words.

What should we conclude about authorship at the 5% significance level?

Parameter: μ = population mean number of new words

- $H_0 : \mu = 8.9$
- $H_1 : \mu > 8.9$
  - can only be $>, <, \neq$

We assume the new author is going to have more words than the original author.

$$\bar{X} \sim N(8.9, \frac{2.5}{\sqrt{6}})$$

R-Code

```
x_bar <- 10.2
mu <- 8.9
sigma_from_sample <- 2.5/sqrt(6)

pnorm(x_bar, mu, sigma_from_sample, lower.tail = FALSE)
[1] 0.1013787
```

Since the outcome is 10 percent, we fail to reject the null hypothesis. We need to default back onto the original assumption. We do not have any definitive truth to accept the null hypothesis but since we have nothing else to fall back onto be need to say its likely to still be true.

Two-sided Tests

Use this when theory does not definitively give an “alternative”.

For Neyman-Pearson tests: critical region is half in left tail and half in right tail of sampling distribution under H_0.

This means that our power will decrease (↓).

For NHST: Find the “one-sided” p-value & double it.