Normal and t Distributions

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

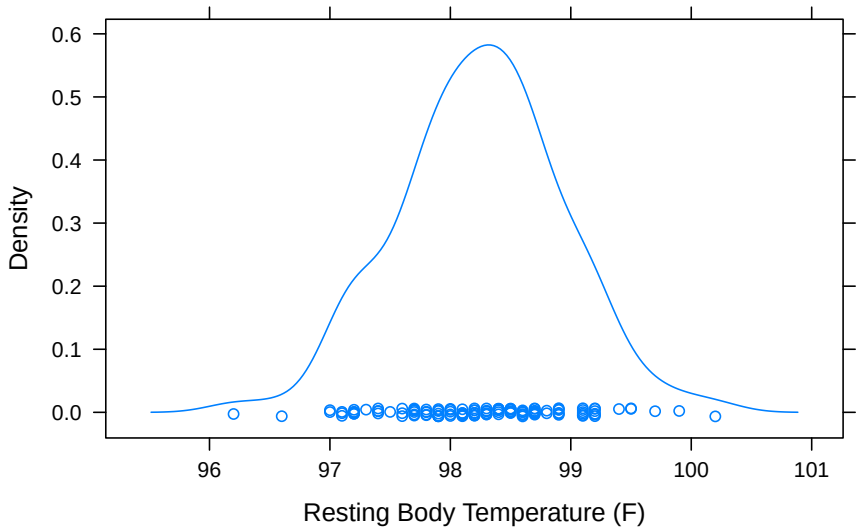
October 11–13, 2011

Case Study

Case Study

Body temperature varies within individuals over time (it can be higher when one is ill with a fever, or during or after physical exertion). However, if we measure the body temperature of a single healthy person when at rest, these measurements vary little from day to day, and we can associate with each person an individual resting body temperture. There is, however, variation among individuals of resting body temperture. A sample of $n = 130$ individuals had an average resting body temperature of 98.25 degrees Fahrenheit and a standard deviation of 0.73 degrees Fahrenheit. The next slide shows an estimated density plot from this sample.

Density Plot

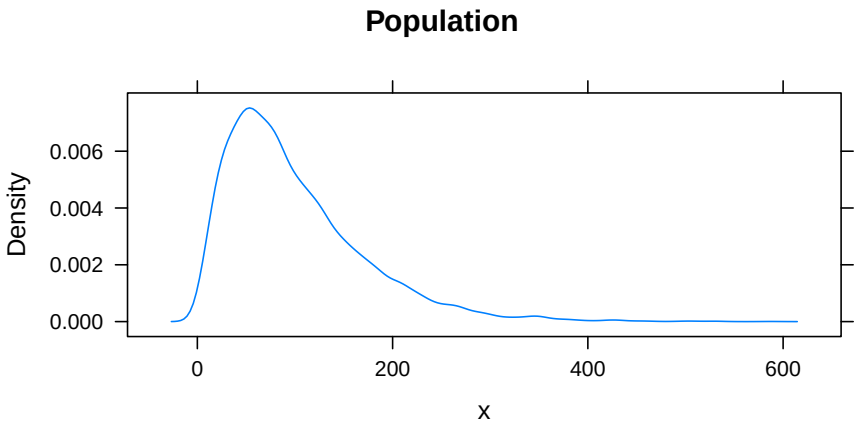


Normal Distributions

- The estimated density has these features:
  - it is bell-shaped;
  - it is nearly symmetric.
- Many (but not all) biological variables have similar shapes.
- One reason is a generalized *the central limit theorem*: random variables that are formed by adding many random effects will be approximately normally distributed.
- Important for inference, even when underlying distributions are not normal, the *sampling distribution of the sample mean* is approximately normal.

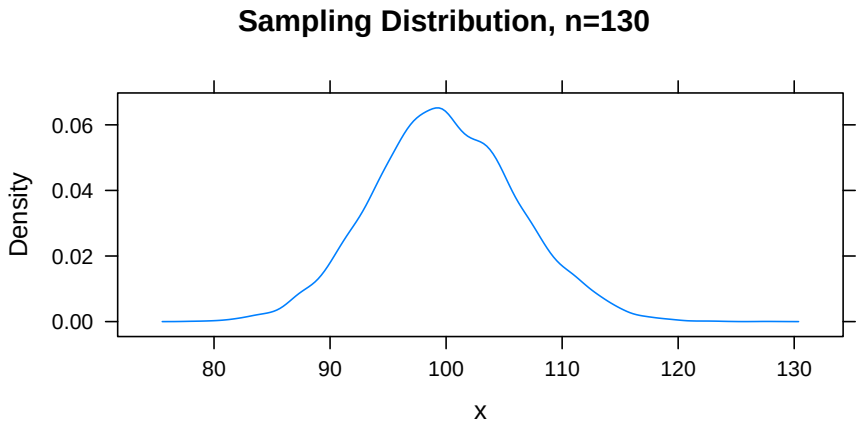
Example: Population

- A population that is skewed.



Example: Sampling Distribution

- Sampling distribution of the sample mean when $n = 130$.



Case Study: Questions

Case Study

- How can we use the sample data to *estimate with confidence* the mean resting body temperture in a population?
- How would we *test the null hypothesis* that the mean resting body temperture in the population is, in fact, equal to the well-known 98.6 degrees Fahrenheit?
- How robust are the methods of inference to nonnormality in the underlying population?
- How large of a sample is needed to ensure that a confidence interval is no larger than some specified amount?

The Big Picture

- Many inference problems with a single *quantitative, continuous variable* may be modeled as a large population (bucket) of individual numbers with a mean μ and standard deviation σ.
- A random sample of size n has a sample mean $\bar{x}$ and sample standard deviation s.
- Inference about μ based on sample data assumes that *the sampling distribution of* $\bar{x}$ is approximately normal with $\mathrm{E}(\bar{x}) = \mu$ and $\mathrm{SD}(\bar{x}) = \sigma/\sqrt{n}$.
- To prepare to understand inference methods for single samples of quantitative data, we need to understand:
  - the normal and related distributions;
  - the sampling distribution of $\bar{x}$.

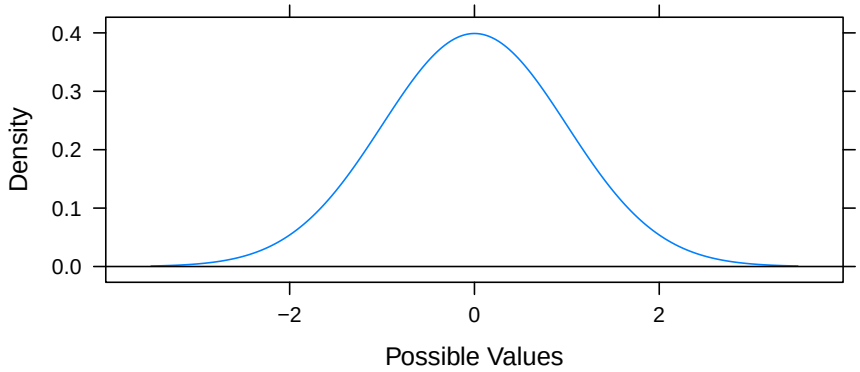
Continuous Distributions

- A *continuous random variable* has possible values over *a continuum*.
- The total probability of one is not in discrete chunks at specific locations, but rather is *ground up like a very fine dust and sprinkled on the number line*.
- We cannot represent the distribution with a table of possible values and the probability of each.
- Instead, we represent the distribution with *a probability density function* which measures the thickness of the probability dust.
- Probability is measured over intervals as the *area under the curve*.
- A legal probability density f:
  - is never negative ($f(x) \geq 0$ for $-\infty < x < \infty$).
  - has a total area under the curve of one ($\int_{-\infty}^{\infty} f(x)dx = 1$).

The Standard Normal Density

- The *standard normal density* is a symmetric, bell-shaped probability density with equation:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad (-\infty < z < \infty)$$



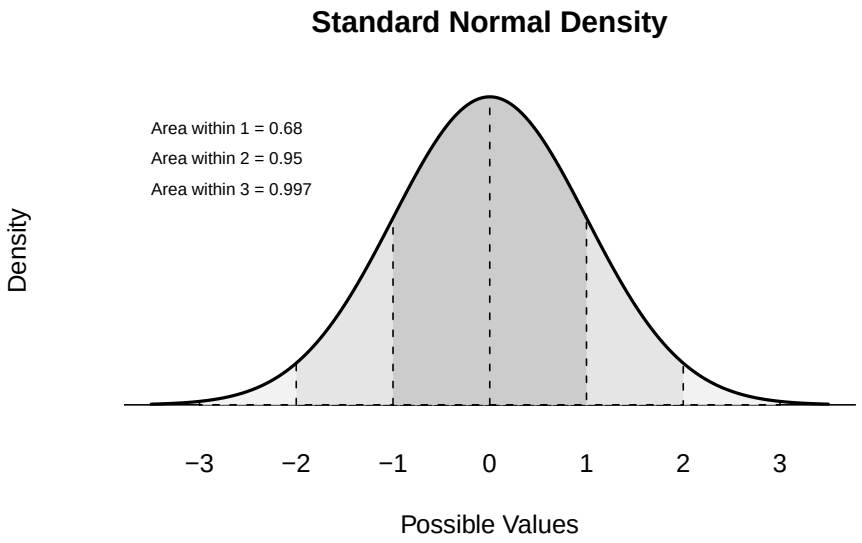
Moments

- The *mean* of the standard normal distribution is $\mu = 0$.
- This point is the center of the density and the point where the density is highest.
- The *standard deviation* of the standard normal distribution is $\sigma = 1$.
- Notice that the points -1 and 1, which are respectively one standard deviation below and above the mean, are at *points of inflection* of the normal curve. (This is useful for roughly estimating the standard deviation from a plotted density or histogram.)

Benchmarks

- The area between -1 and 1 under a standard normal curve is approximately 68%.
- The area between -2 and 2 under a standard normal curve is approximately 95%.
- More precisely, the area between -1.96 and $1.96 \doteq 0.9500$, which is why we have used 1.96 for 95% confidence intervals for proportions.

Standard Normal Density



General Areas

- There is no formula to calculate general areas under the standard normal curve.
- (The integral of the density has no closed form solution.)
- We prefer to use R to find probabilities.
- You also need to learn to use normal tables for exams.

R

- The function `pnorm()` calculates probabilities under the standard normal curve by finding *the area to the left*.
- For example, the area to the left of -1.57 is

```
> pnorm(-1.57)
[1] 0.05820756
```

and the area to the right of 2.12 is

```
> 1 - pnorm(2.12)
[1] 0.01700302
```


Tables

- The table on pages 672–673 displays right tail probabilities for $z = 0$ to $z = 4.09$.
- A point on the axis rounded to two decimal places $a.bc$ corresponds to a row for $a.b$ and a column for c.
- The number in the table for this row and column is the area to the right.
- Symmetry of the normal curve and the fact that the total area is one are needed.
- The area to the left of -1.57 is the area to the right of 1.57 which is 0.05821 in the table.
- The area to the right of 2.12 is 0.01711.
- When using the table, it is best to *draw a rough sketch of the curve and shade in the desired area*. This practice allows one to approximate the correct probability and catch simple errors.
- Find the area between $z = -1.64$ and $z = 2.55$ on the board.

R

- The function `qnorm()` is the inverse of `pnorm()` and finds a *quantile*, or location where a given area is to the right.
- For example, the 0.9 quantile of the standard normal curve is

  ```
  > qnorm(0.9)
  [1] 1.281552
  ```

  and the number z so that the area between $-z$ and z is 0.99 is

  ```
  > qnorm(0.995)
  [1] 2.575829
  ```

  since the area to the left of $-z$ and to the right of z must each be $(1 - 0.99)/2 = 0.005$ and $1 - 0.005 = 0.995$.
- Draw a sketch!

Tables

- Finding quantiles from the normal table almost always requires some round off error.
- To find the number z so that the area between $-z$ and z is 0.99 requires finding the probability 0.00500 in the middle of the table.
- We see $z = 2.57$ has a right tail area of 0.00508 and $z = 2.58$ has a right ail area of 0.00494, so the value of z we seek is between 2.57 and 2.58.
- For exam purposes, it is okay to pick the closest, here 2.57.
- Use the table to find the 0.03 quantile as accurately as possible.
- Draw a sketch!

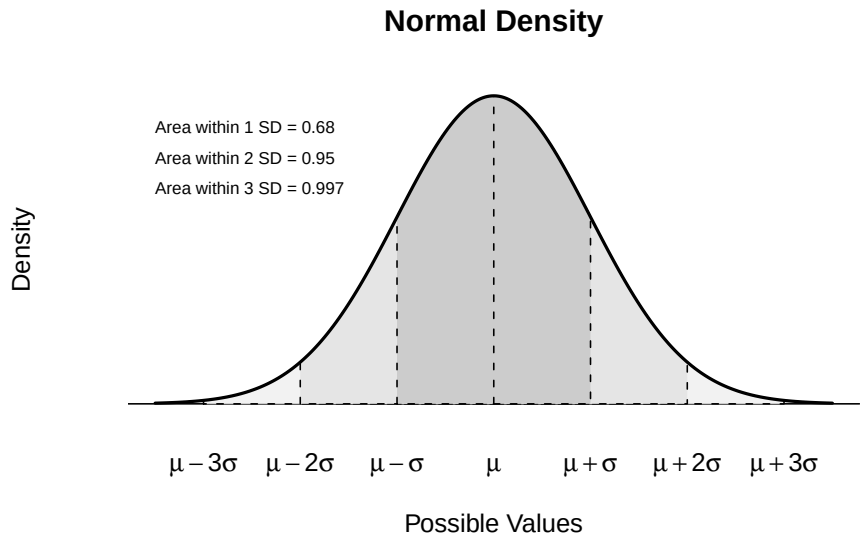
General Normal Density

- The *general normal density* with mean μ and standard deviation σ is a symmetric, bell-shaped probability density with equation:

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

- Sketches of general normal curves have the same shape as standard normal curves, but have rescaled axes.

General Normal Density



All Normal Curves Have the Same Shape

- All normal curves have the same shape, and are simply rescaled versions of the standard normal density.
- Consequently, every area under a general normal curve corresponds to an area under the standard normal curve.
- The key *standardization formula* is

$$z = \frac{x - \mu}{\sigma}$$

- Solving for x yields

$$x = \mu + z\sigma$$

which says algebraically that x is z standard deviations above the mean.

Normal Tail Probability

Example

- If $X \sim \mathrm{N}(100, 2)$, find $\mathrm{P}(X > 97.5)$.
- Solution:

$$\begin{aligned} \mathrm{P}(X > 97.5) &= \mathrm{P}\left(\frac{X - 100}{2} > \frac{97.5 - 100}{2}\right) \\ &= \mathrm{P}(Z > -1.25) \\ &= 1 - \mathrm{P}(Z > 1.25) \\ &= 0.8944 \end{aligned}$$

Normal Quantiles

Example

- If $X \sim \mathrm{N}(100, 2)$, find the cutoff values for the middle 70% of the distribution.
- Solution: The cutoff points will be the 0.15 and 0.85 quantiles.
- From the table, $1.03 < z < 1.04$ and $z = 1.04$ is closest.
- Thus, the cutoff points are the mean plus or minus 1.04 standard deviations.

$$100 - 1.04(2) = 97.92, \quad 100 + 1.04(2) = 102.08$$

- In R, a single call to `qnorm()` finds these cutoffs.

```
> qnorm(c(0.15, 0.85), 100, 2)
[1]  97.92713 102.07287
```


Case Study

Example

In a population, suppose that:

- the mean resting body temperature is 98.25 degrees Fahrenheit;
- the standard deviation is 0.73 degrees Fahrenheit;
- resting body temperatures are normally distributed.

Let X be the resting body temperature of a randomly chosen individual. Find:

1. $\mathrm{P}(X < 98)$, the proportion of individuals with temperature less than 98.
2. $\mathrm{P}(98 < X < 100)$, the proportion of individuals with temperature between 98 and 100.
3. The 0.90 quantile of the distribution.
4. The cutoff values for the middle 50% of the distribution.

Answers (with R, table will be close)

1. 0.366
2. 0.6257
3. 99.19
4. 97.76 and 98.74

The χ^2 Distribution

- The χ^2 distribution is used to find p-values for the test of independence and the G-test we saw earlier for contingency tables.
- Now that the normal distribution has been introduced, we can better motivate the χ^2 distribution.

Definition

If $Z_1, \ldots, Z_k$ are independent standard normal random variables, then

$$X^2 = Z_1^2 + \cdots + Z_k^2$$

has a χ^2 distribution with k degrees of freedom.

The χ^2 Distribution

- The functions `pchisq()` and `qchisq()` find probabilities and quantiles, respectively, from the χ^2 distributions.
- The table on pages 669–671 has the same information for limited numbers of quantiles for each χ^2 distribution with 100 or fewer degrees of freedom.
- Unlike the normal distributions where all normal curves are just rescalings of the standard normal curve, each χ^2 distribution is different.

t Distribution

Definition

If Z is a standard normal random variable and if X^2 is a χ^2 random variable with k degrees of freedom, then

$$T = \frac{Z}{\sqrt{X^2/k}}$$

has a t distribution with k degrees of freedom.

- t densities are symmetric, bell-shaped, and centered at 0 just like the standard normal density, but are more spread out (higher variance).
- As the degrees of freedom increases, the t distributions converge to the standard normal.
- t distributions will be useful for statistical inference for one or more populations of quantitative variables.

The Central Limit Theorem

The Central Limit Theorem

If $X_1, \ldots, X_n$ are an independent sample from a common distribution F with mean $\mathrm{E}(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$, (which need not be normal), then

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

is approximately normal with $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ if the sample size n is sufficiently large.

- The central limit theorem (and its cousins) justifies almost all inference methods the rest of the semester.

Mean of the Sampling Distribution of $\bar{X}$

- The mean of the sampling distribution of $\bar{X}$ is found using the linearity properties of expectation.

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{\sum_{i=1}^n X_i}{n}\right) \\
&= \left(\frac{1}{n}\right)\mathrm{E}(X_1 + \cdots + X_n) \\
&= \left(\frac{1}{n}\right)(\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)) \\
&= \left(\frac{1}{n}\right) n\mu \\
&= \mu
\end{aligned}$$

Variance of the Sampling Distribution of $\bar{X}$

- The variance of the sampling distribution of $\bar{X}$ is found using the properties of variances of sums.

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i}{n}\right) \\
&= \left(\frac{1}{n}\right)^2 \mathrm{Var}(X_1 + \cdots + X_n) \\
&= \left(\frac{1}{n}\right)^2 (\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) \\
&= \left(\frac{1}{n}\right)^2 n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Also, $\mathrm{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

Case Study

Example

In a population, suppose that:

- the mean resting body temperature is 98.25 degrees Fahrenheit;
- the standard deviation is 0.73 degrees Fahrenheit;
- resting body temperatures are normally distributed.

Let $X_1, \ldots, X_{40}$ be the resting body temperatures of 40 randomly chosen individuals from the population. Find:

1. $\mathrm{P}(\bar{X} < 98)$, the probability that the sample mean is less than 98.
2. $\mathrm{P}(98 < \bar{X} < 100)$, the probability that the sample mean is between 98 and 100.
3. the 0.90 quantile of the sampling distribution of $\bar{X}$.
4. The cutoff values for the middle 50% of the sampling distribution of $\bar{X}$.

Answers (with R, table will be close)

1. 0.0152
2. 0.9848
3. 98.4
4. 98.17 and 98.33