Confidence Intervals, normal and t-distribution

Sumanta Basu

Reading

Reading: Textbook sections 4.2, 4.4, 4.5.1, 5.1.1 - 5.1.4

Recommended Reading: Sections 3.2 and 3.3 of Chapter 3 supplement (on blackboard)

Recommended Exercise: 4.7, 4.9a,b, 4.11, 4.12, 4.13

What we’ve observed in Cherry Blossom Race Data

Let $\bar{X}_n$ is the sample mean of a SRS of size n from a population with mean μ and standard deviation σ. Then:

$$E(\bar{X}_n) = \mu$$

$$\mathrm{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

And what's more…

When n is "large enough", $\bar{X}_n$ is approximately normal!!

$$\bar{X}_n \text{ is approximately } N(\mu, \sigma/\sqrt{n})$$

Central Limit Theorem

> If $X_1, \ldots, X_n$ are independent draws from a distribution with mean μ and standard deviation σ, then for large n, the sample mean $\bar{X}_n$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- remarkable since individual X_i 's don't have to look at all like a normal distribution

- how large should n be? Depends, but if distribution of X_i 's is not strongly skewed, say $n \geq 30$

An Important Special Case: Bernoulli

Suppose $X_1, \ldots, X_n$ are independent coin flips, i.e., $X_i \sim \text{Bernoulli}(p)$.

The sample proportion, sometimes written $\hat{p}_n$, is just $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Recall $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$.

CLT tells us

$$\hat{p}_n \approx N(p, \sqrt{p(1-p)/n})$$

> Rule of thumb: Used to create confidence interval for p when $n\hat{p}_n \geq 10$ and $n(1-\hat{p}_n) \geq 10$

Inference

Sampling distribution: A probabilistic description of how the observed values of a numerical summary statistic (e.g., sample mean) behave under repeated SRS.

This concept underlies all basic statistical inference procedures – its importance cannot be overstated!

In practice: we only collect one sample.

Question: how can we combine the information from a single SRS about a population parameter with our knowledge of sampling distributions in order to perform statistical inference?

Two primary goals

1. A **confidence interval** - a range of plausible values for a (population) parameter, based on the data obtained from our observed sample.

2. A **hypothesis (or significance) test** - an assessment of whether the observed value of a statistic computed using the sample data is consistent with or divergent from some hypothesized value of the (population) parameter.

Note: these get at *"what is μ?"* better than just reporting a single **point estimate** (e.g., $\bar{x} = 33.8$)

Five Examples we will see in class

Population mean (μ): *average score in a class*

Population proportion (p): *proportion of students with grades* $A-$ *or better*

Difference between two population proportions ($p_1 - p_2$): *Compare proportion of students with* $A-$ *or better across two labs*

Mean difference between two unrelated populations ($\mu_1 - \mu_2$): *Compare average scores across different labs*

Mean difference between two related populations (μ_d): *Compare average scores of Prelim 1 and Prelim 2*

Five Examples we will see in class

Population mean (μ): *average score in a class*

Population proportion (p): *proportion of students with grades $A-$ or better*

Difference between two population proportions ($p_1 - p_2$): *Compare proportion of students with $A-$ or better across two labs*

Mean difference between two unrelated populations ($\mu_1 - \mu_2$): *Compare average scores across different labs*

Mean difference between two related populations (μ_d): *Compare average scores of Prelim 1 and Prelim 2*

Confidence Interval (CI)

… Contd. from last lecture

The random interval $[\bar{X}_n - 3.2, \bar{X}_n + 3.2]$ is called a 96% **confidence interval**. Here, 96% is said to be its **confidence level**.

$$P([\bar{X}_n - 3.2, \bar{X}_n + 3.2] \text{ includes } \mu) = 96\%$$

Note

In practice, we only observe one realization of the random variable $\bar{X}_n$, say, $\bar{x} = 33.8$.

In this case, the realization of our confidence interval is $[30.6, 37.0]$.

We'll refer to $[30.6, 37.0]$ as a **96% confidence interval for μ** even though we technically should call it a *realization* of a 96% confidence interval.

Can we choose the width to get a desired confidence level?

Want to choose w so that

$$P(\bar{X}_n - w \leq \mu \leq \bar{X}_n + w) = 95\%$$

(or some other probability)

If only we knew the distribution of $\bar{X}_n$.

Flashback - CLT

If $X_1, \ldots, X_n$ are independent draws from a distribution with mean μ and standard deviation σ, then for large n, the sample mean $\bar{X}_n$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$:

$$\bar{X}_n \approx N(\mu, \sigma/\sqrt{n})$$

Also recall: a normal random variable falls within *2 standard deviations of its mean* with probability about 95%.

$$P\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) \approx 95\%$$

Flash forward - Can we choose the width to get a desired confidence level?

Want to choose w so that

$$P(\bar{X}_n - w \leq \mu \leq \bar{X}_n + w) = 95\%$$

(or some other probability)

If only we knew the distribution of $\bar{X}_n$.

Flash forward - Can we choose the width to get a desired confidence level?

Want to choose w so that

$$P(\bar{X}_n - w \leq \mu \leq \bar{X}_n + w) = 95\%$$

(or some other probability)

If only we knew the distribution of $\bar{X}_n$.

$$P\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) \approx 95\%$$

is equivalent to (after some simple algebra)

$$P\left(\bar{X}_n - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{2\sigma}{\sqrt{n}}\right) \approx 95\%$$

What we've done

Using the CLT, we have constructed an **approximate 95% confidence interval for μ**!

$$\left[\bar{X}_n - \frac{2\sigma}{\sqrt{n}}, \bar{X}_n + \frac{2\sigma}{\sqrt{n}}\right]$$

Why is it *approximate*?

- CLT (better approx for n large and X_i 's distribution not too skewed)
- 2 standard deviations rule of thumb

In example, $2\sigma/\sqrt{n} = 3.2$. *Approximation actually worked!*

Is there something unrealistic about this situation?

$$\left[\bar{X}_n - \frac{2\sigma}{\sqrt{n}}, \bar{X}_n + \frac{2\sigma}{\sqrt{n}}\right]$$

- We have assumed that (population) variance, σ^2, of the X_i's distribution is known.
- Seems unlikely we'd know variance of age of random runner selected but not the mean.
- This was a simplifying assumption… in practice we will want to estimate σ and still get a confidence interval for μ.

General confidence levels

Suppose we construct a **confidence interval of level $100(1-\alpha)\%$**.

That is, *a method that computes an interval based on a sample such that, imagining repeated sampling, $100(1-\alpha)\%$ of such intervals would succeed in including the population parameter μ.*

General confidence levels

Recall

$$\bar{X}_n \approx N(\mu, \sigma/\sqrt{n}) \text{ is equivalent to } \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

General confidence levels

Recall

$$\bar{X}_n \approx N(\mu, \sigma/\sqrt{n}) \text{ is equivalent to } \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

Getting a quantile

To get $\alpha/2$ quantile of $N(0, 1)$, use `qnorm(alpha/2)`

For a 95% confidence interval…

```r
qnorm(0.025)
```

```
## [1] -1.959964
```

so we use 1.96 (hence our rule of thumb of 2)

Example

Suppose we know $\sigma = 10.16$ years and our sample is

```
##  [1] 31 30 48 41 30 33 25 28 33 39
```

```
## [1] 33.8
```

Compute a 99% confidence interval.

Example

What is α? Draw a picture!

```
alpha = 0.01; sigma = 10.16; n = 10
xbar = mean(x1)
zvalue = -qnorm(alpha/2)
xbar - zvalue * sigma / sqrt(n) # lower
```

```
## [1] 25.52418
```

```
xbar + zvalue * sigma / sqrt(n) # upper
```

```
## [1] 42.07582
```

In words

"An approximate 99% confidence interval for the expected age of someone finishing the race is [25.5,42.1]."

"We are 99% confident that the average age of a runner completing the race is between 25.5 and 42.1."

What this actually means:

If we repeatedly gathered SRS's of size 10 and computed an interval in this manner each time, then in the long run 99% of these intervals would include the (population) mean age of someone finishing the race.

Question

Is the following correct?

$$P(25.5 \leq \mu \leq 42.1) \approx 99\%$$

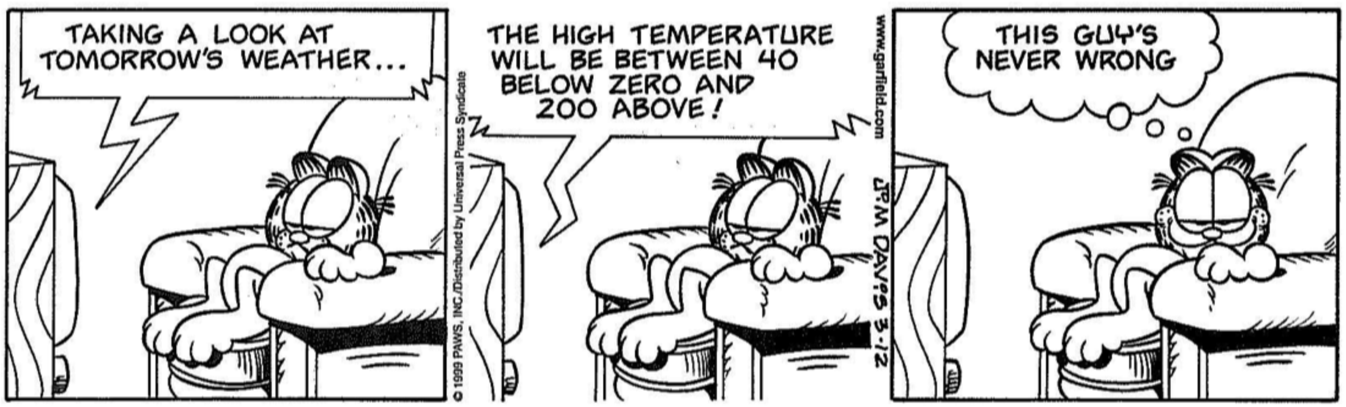
No. The numbers 25.5, 42.1, and μ are not random. So either it holds or it doesn't (with probability 1)

What is true:

$$P(\bar{X}_n - 2.58\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + 2.58\sigma/\sqrt{n}) \approx 99\%$$

$\bar{X}_n$ is random, so it makes sense to talk about probability. Not so with $\bar{x}_n$ (a realization of $\bar{X}_n$).

Is a level 100% confidence interval useful?



When can we treat sample mean as normal?

1. **When CLT applies** - independent observations, $n > 30$, data distribution not strongly skewed
2. **When data distribution is itself nearly normal** - independent observations (in this case, we don't need n large)

In both cases,

$$\bar{X}_n \approx N(\mu, \sigma/\sqrt{n})$$

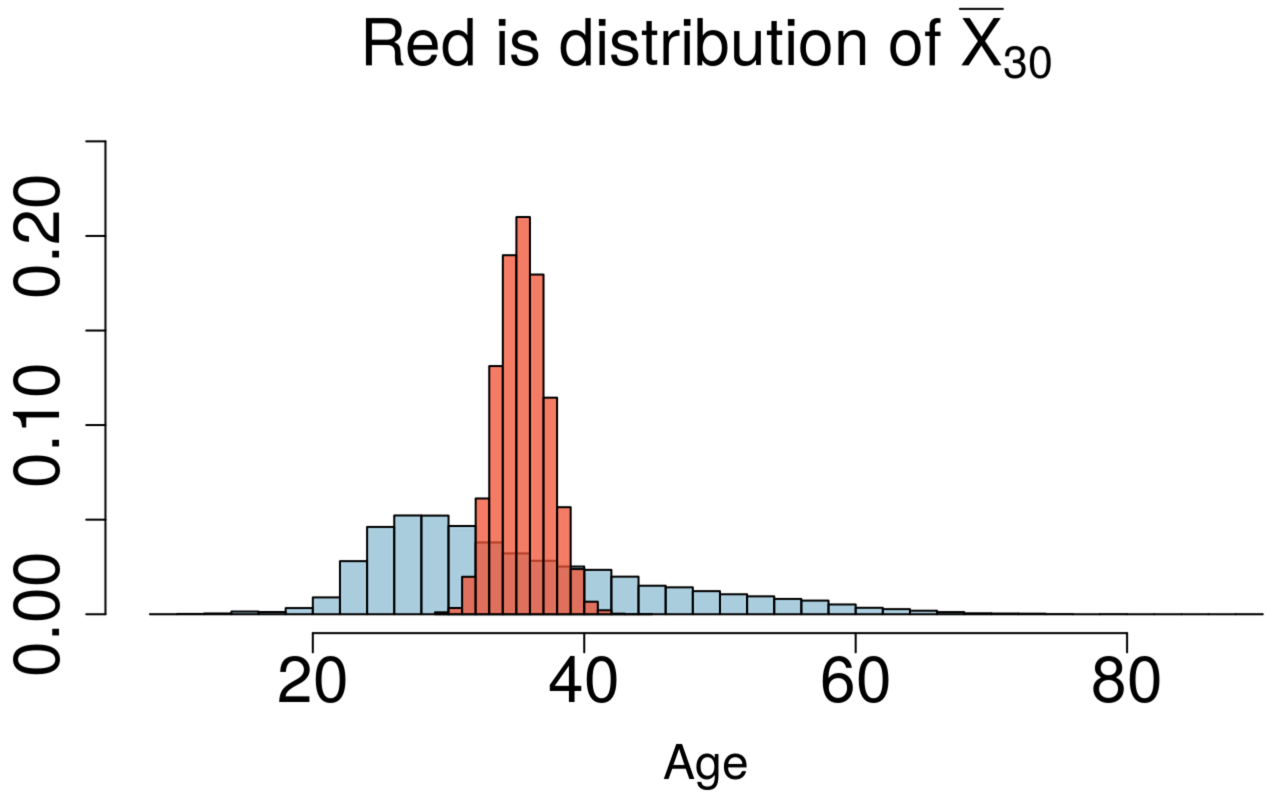
that is

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

write on board

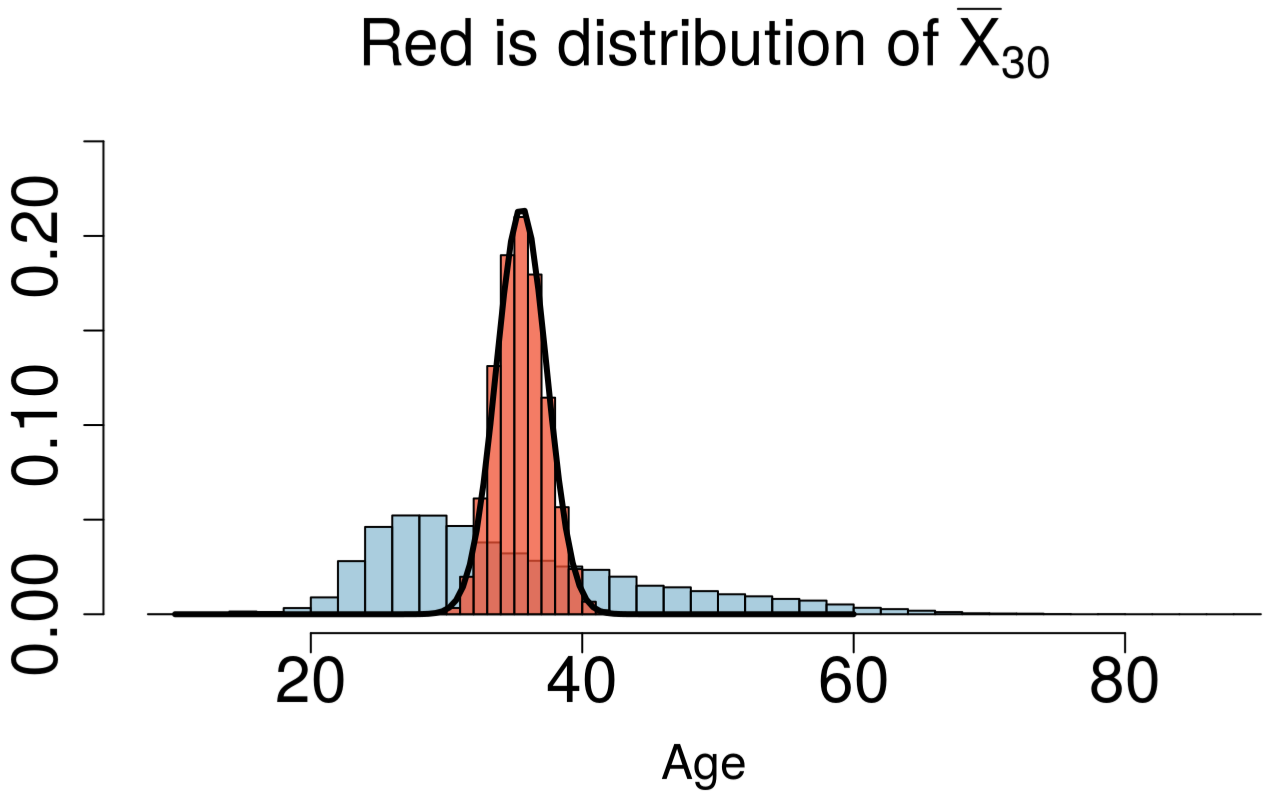
Case 1: When CLT applies

Runner data is skewed, but $\bar{X}_n$ should be normal for $n > 30$



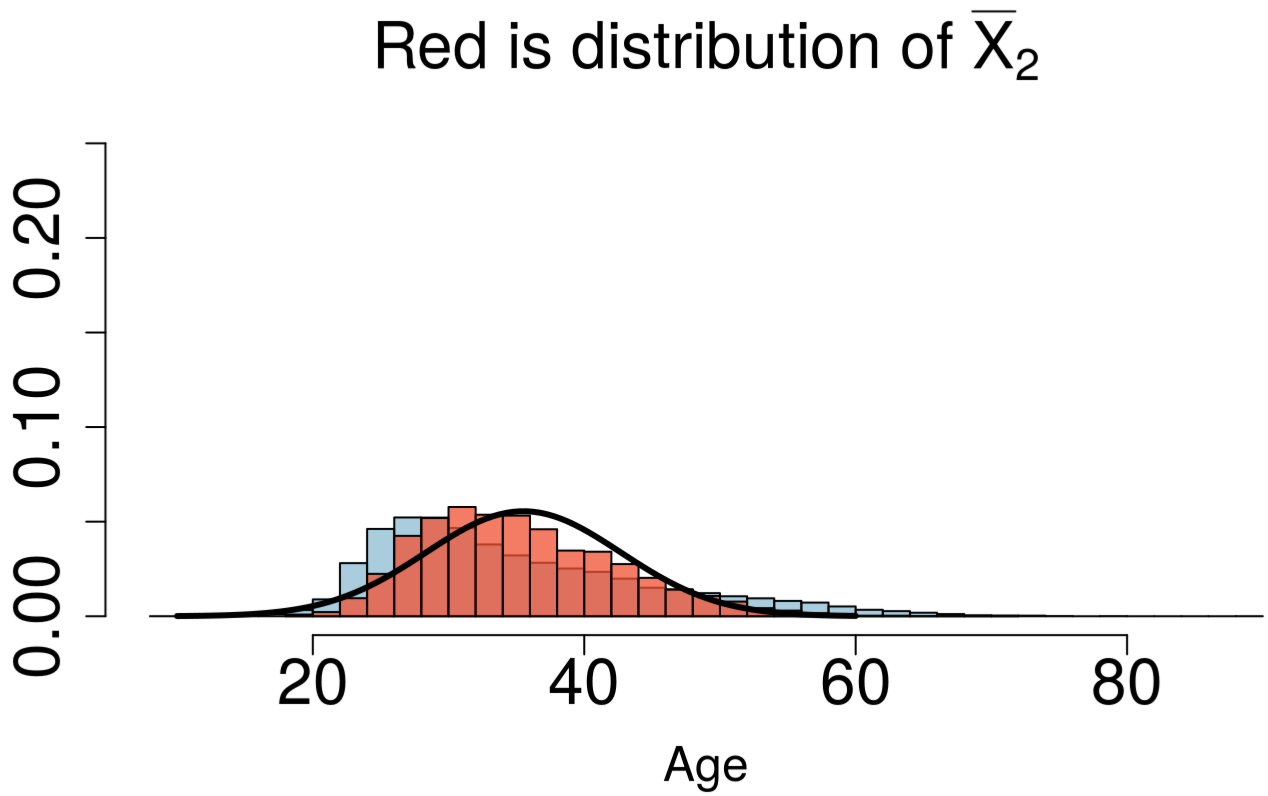
Case 1: When CLT applies

Runner data is skewed, but $\bar{X}_n$ should be normal for $n > 30$



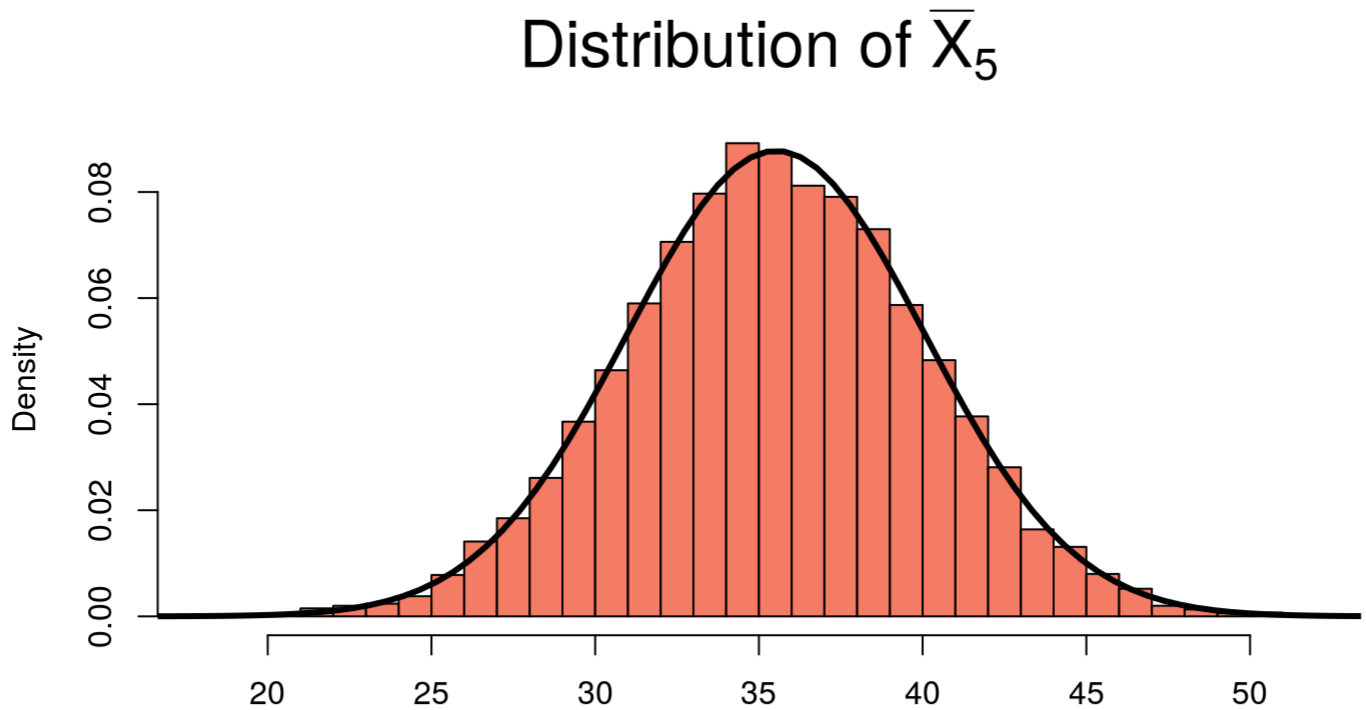
Not Case 1: When CLT does not apply

Runner data is skewed, $\bar{X}_n$ does not look $N(\mu, \sigma/\sqrt{n})$ for $n = 2$



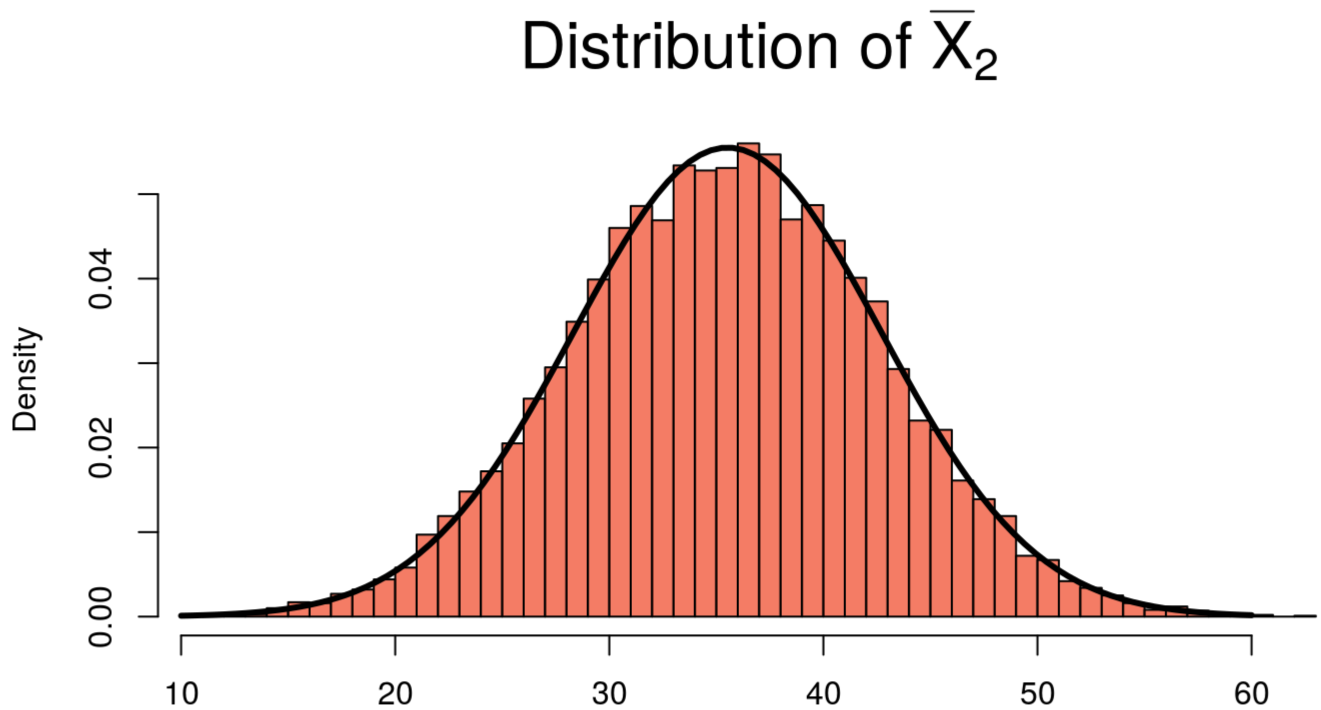
Case 2: Nearly normal data (n=5)

Draw $X_1, \ldots, X_5 \sim N(\mu, \sigma)$. See $\bar{X}_n \sim N(\mu, \sigma/\sqrt{n})$



Case 2: Nearly normal data (n=2)

Draw $X_1, X_2 \sim N(\mu, \sigma)$. See $\bar{X}_n \sim N(\mu, \sigma/\sqrt{n})$



Logic for getting confidence intervals

Used

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

plus a bit of algebra to get

$$P(\bar{X}_n - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\sigma/\sqrt{n}) \approx 1 - \alpha$$

What if σ is unknown?

Idea: replace σ by $S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$

But does our logic still hold?

Dealing with unknown variance

Would like to simply say that instead of

$$P(\bar{X}_n - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\sigma/\sqrt{n}) \approx 1 - \alpha$$

we have

$$P(\bar{X}_n - z_{\alpha/2}S_n/\sqrt{n} \leq \mu \leq \bar{X}_n + z_{\alpha/2}S_n/\sqrt{n}) \approx 1 - \alpha$$

However, there's a problem:

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \not\approx N(0, 1)$$

so shouldn't be using $z_{\alpha/2}$, which is quantile from $N(0, 1)$.

Why is it not normal?

Intuitively,

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \qquad \text{(let's call this } T_n\text{)}$$

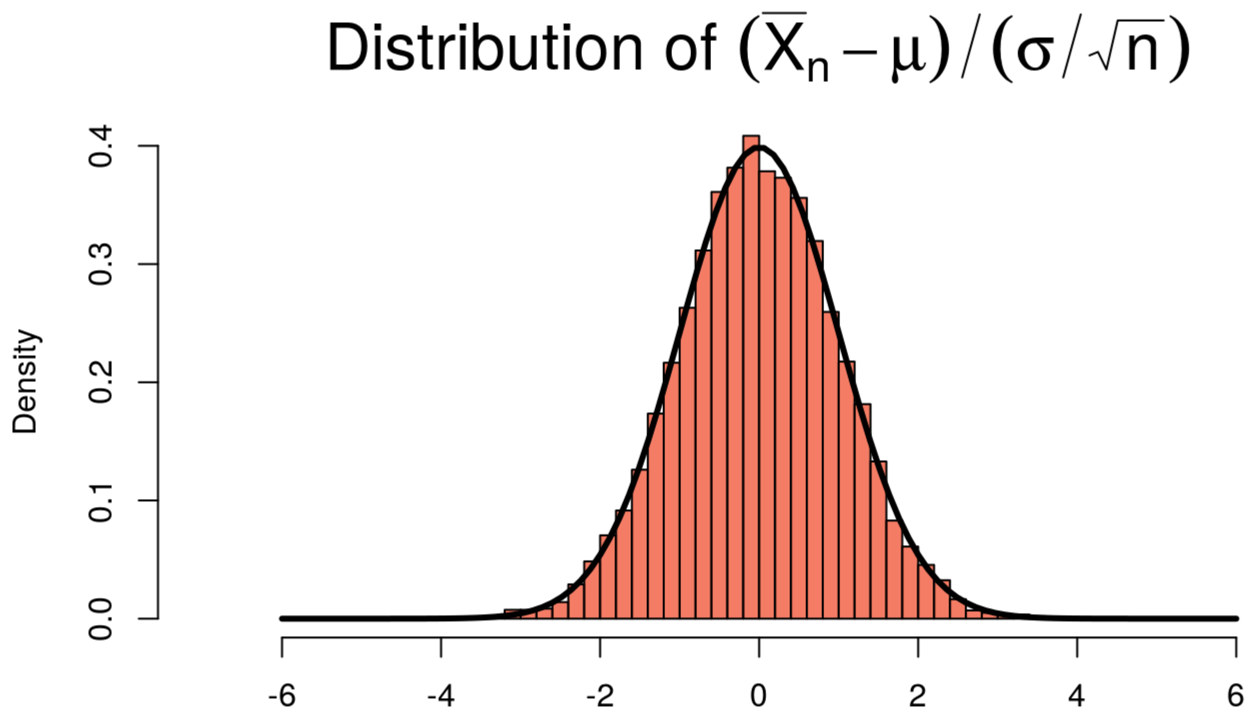
has more variability “in it” than

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

since S_n is also random.

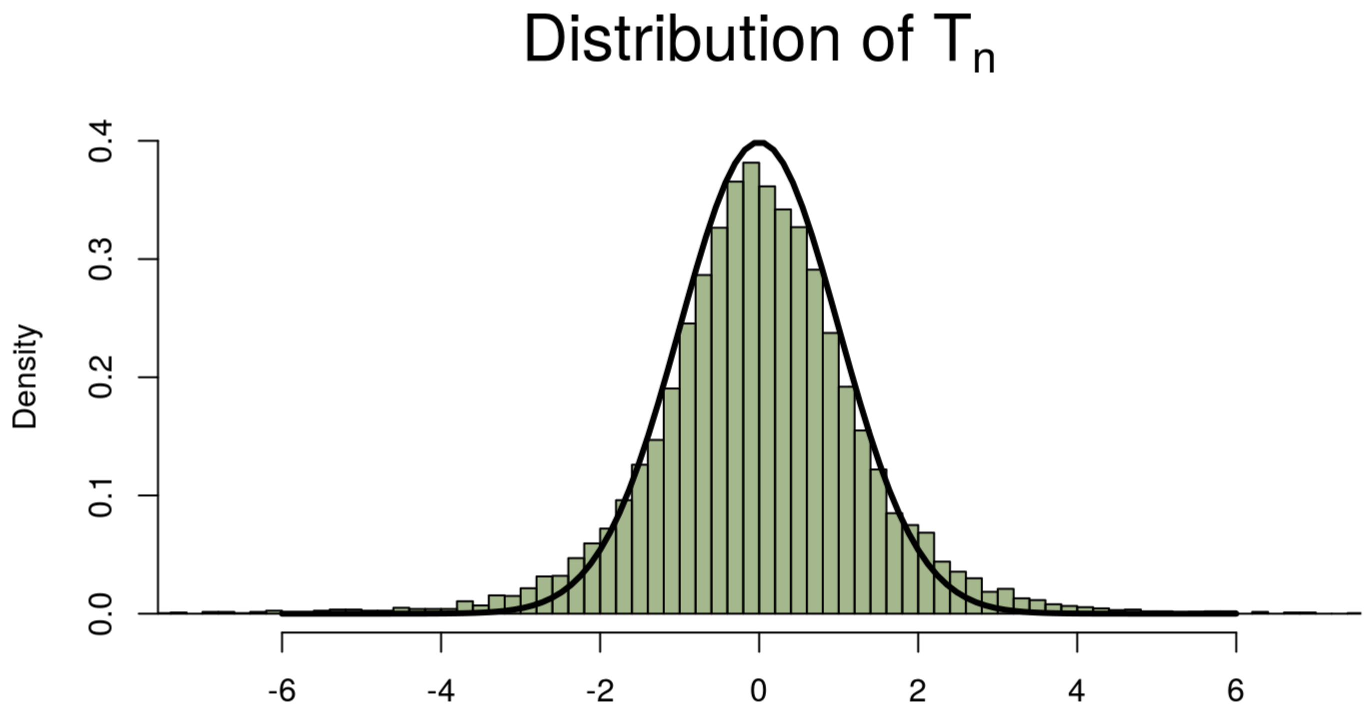
Monte Carlo simulation (normal data, n=5)

Draw $X_1, \ldots, X_5 \sim N(\mu, \sigma)$: See $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$



Monte Carlo simulation (normal data, n=5)

Same as before. See $T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ has heavier tails than $N(0, 1)$



Student's t-distribution

If $X_1, \ldots, X_n \sim N(\mu, \sigma)$ are independent, then

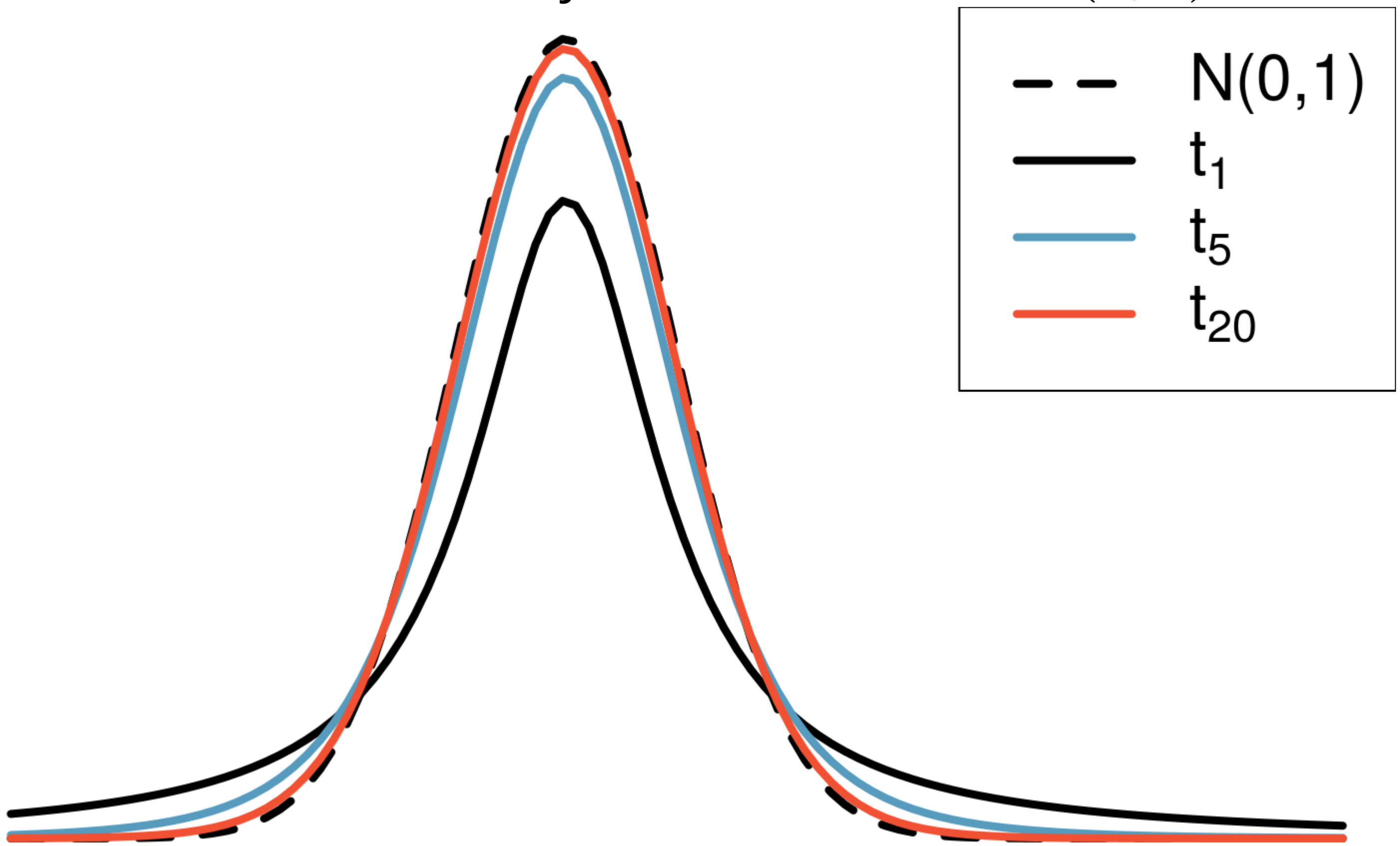
$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

In words, we say that T_n has a **t-distribution with $n-1$ degrees of freedom**.

t_n denotes this distribution.

Student's t-distribution

For small n has noticeably heavier tails than $N(0, 1)$.



William Gosset’s 1908 paper

William Gosset’s 1908 paper

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

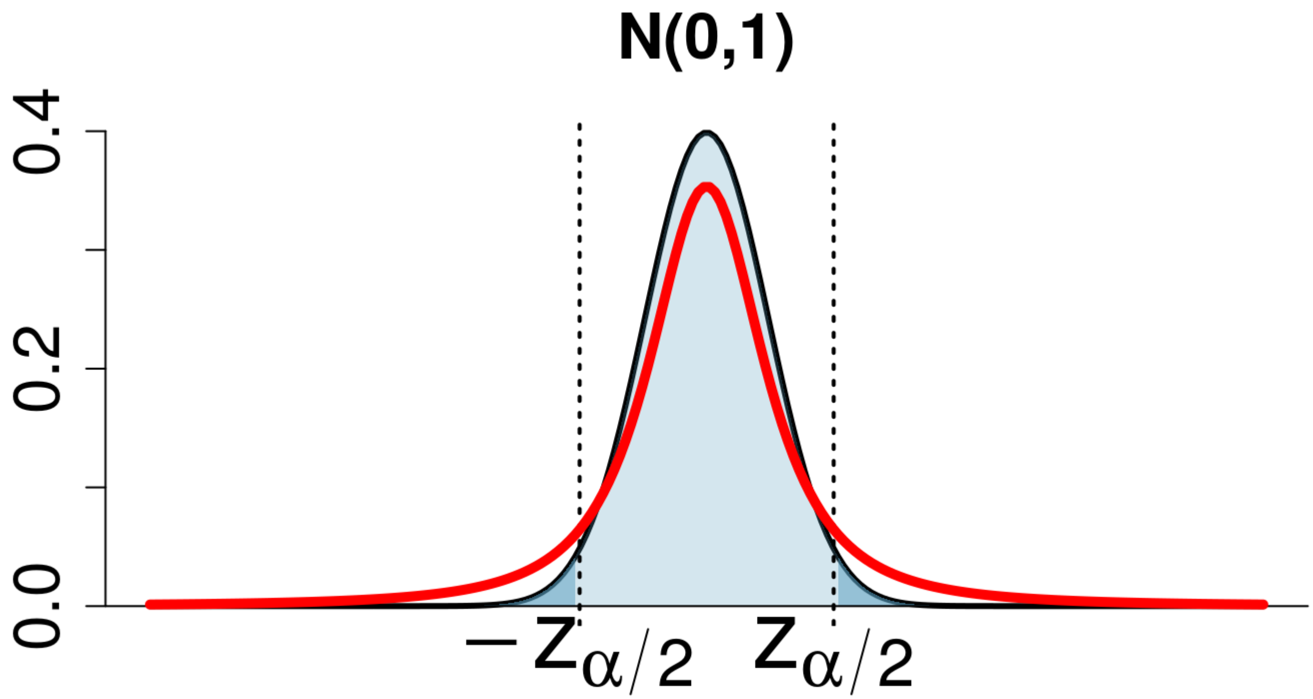
BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a “population” of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

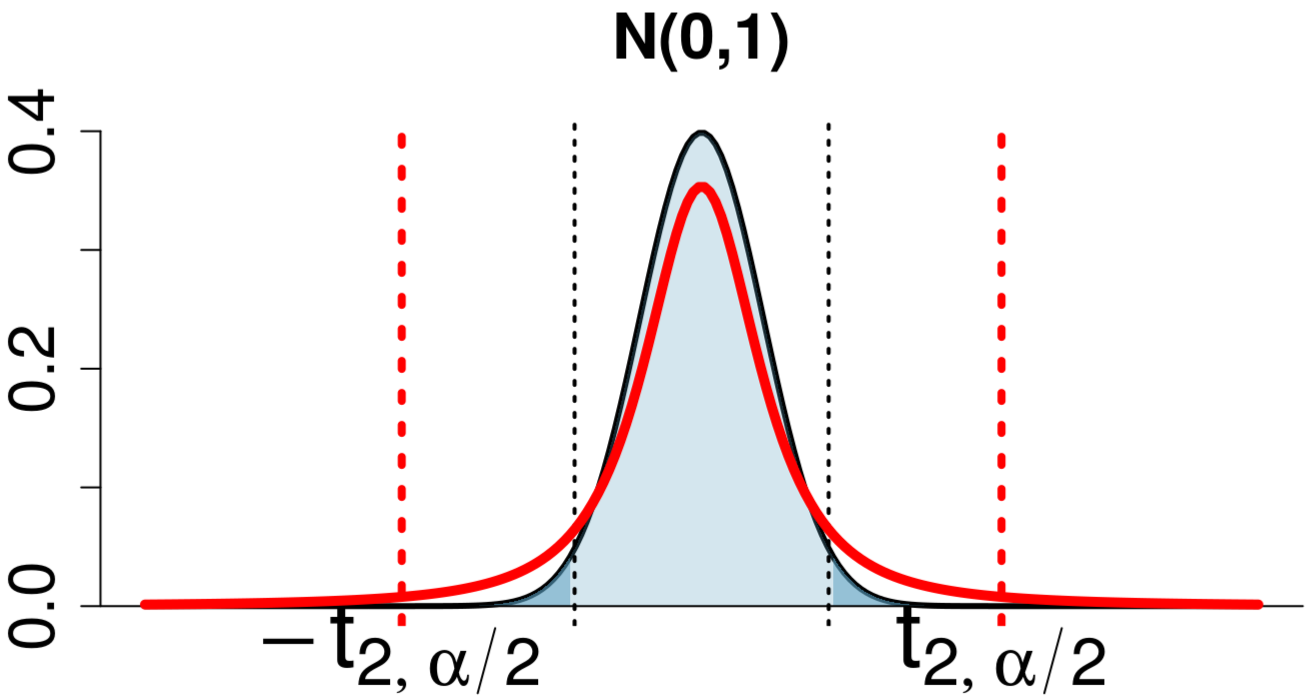
Quantiles of t-distribution

Where is the $1 - \alpha/2$ quantile of t_{df} (which we'll call $t_{df,\alpha/2}$)?



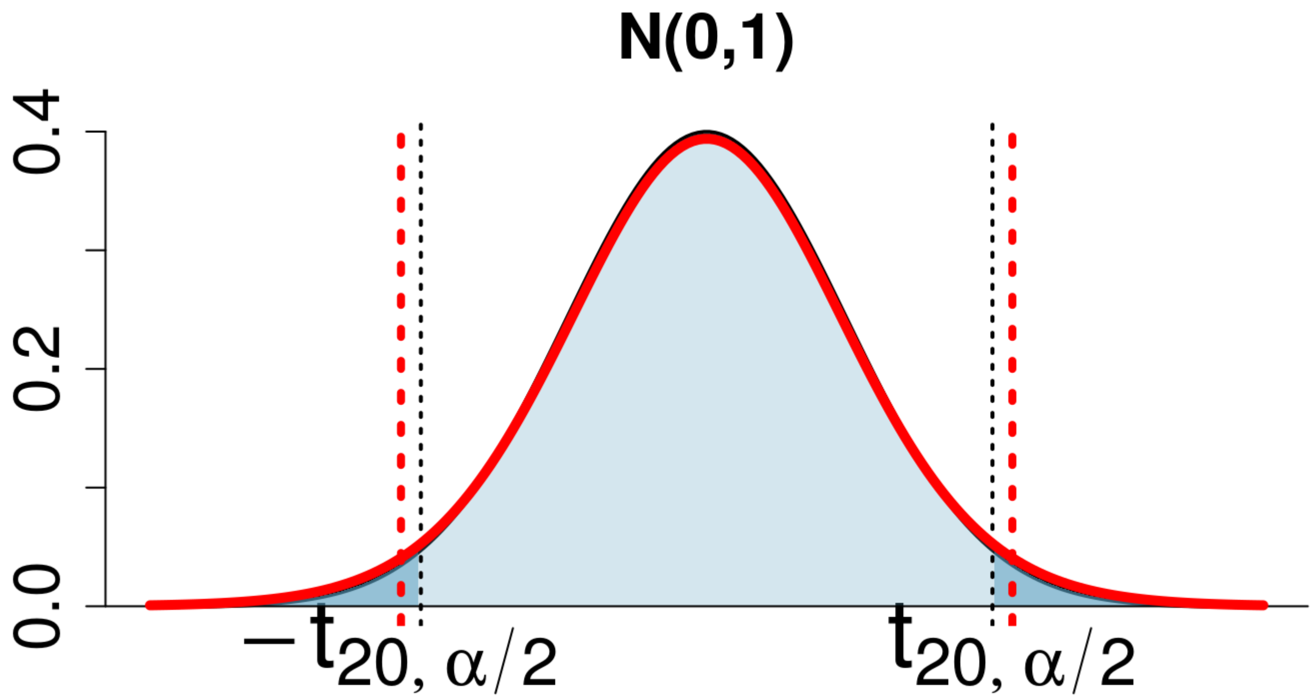
Quantiles of t-distribution

Where is the $1-\alpha/2$ quantile of t_{df} (which we'll call $t_{df,\alpha/2}$)?



Quantiles of t-distribution

Where is the $1 - \alpha/2$ quantile of t_{df} (which we'll call $t_{df,\alpha/2}$)?



Using t-distribution for confidence interval

If $X_1, \ldots, X_n$ are roughly $N(\mu, \sigma)$ and independent, then

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

which (by identical algebra as before) means

$$P(\bar{X}_n - t_{n-1,\alpha/2} S_n/\sqrt{n} \leq \mu \leq \bar{X}_n + t_{n-1,\alpha/2} S_n/\sqrt{n}) \approx 1 - \alpha$$

Recap: t-distribution for a CI

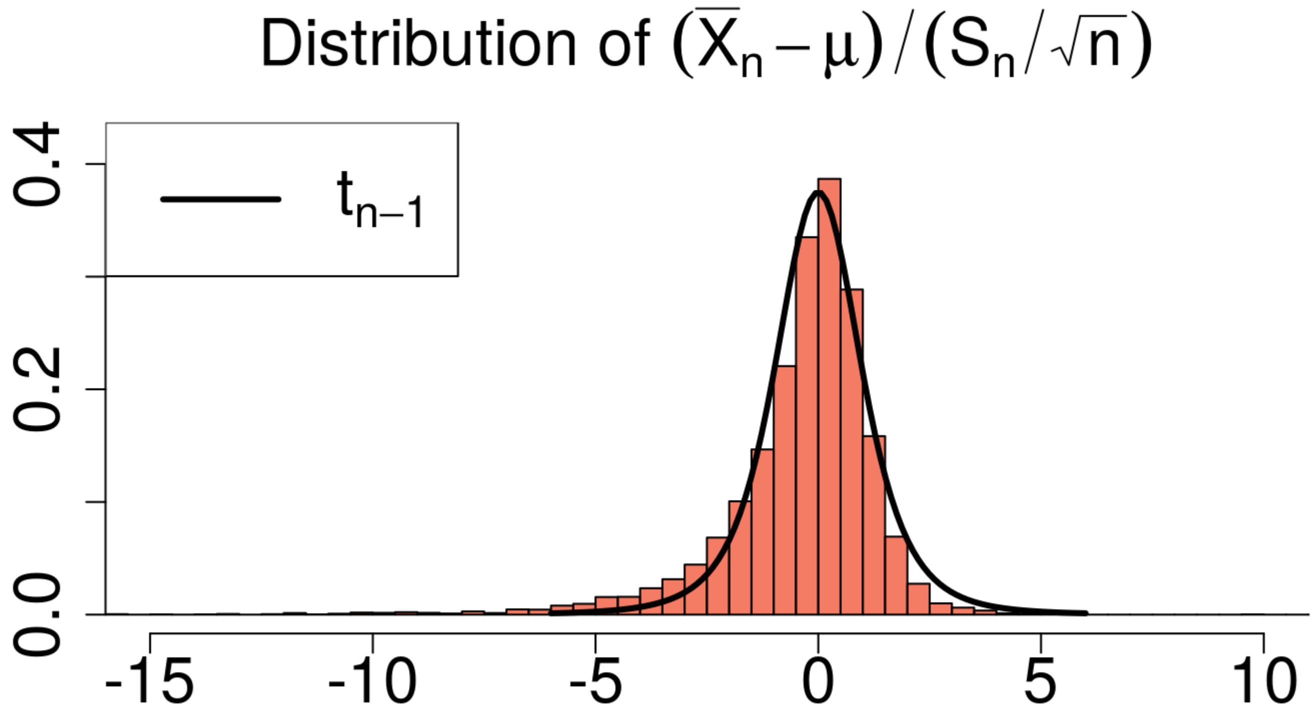
1. Use the **t-distribution** when…

- observations are independent
- data's distribution is nearly normal
- σ is unknown

1. When n is large (say, $n > 30$), t_{n-1} is so close to $N(0, 1)$ that it doesn't make much of a difference whether you use $t_{n-1,\alpha/2}$ versus $z_{\alpha/2}$.

Intuition: when $n > 30$, using S_n is just about the same as using σ.

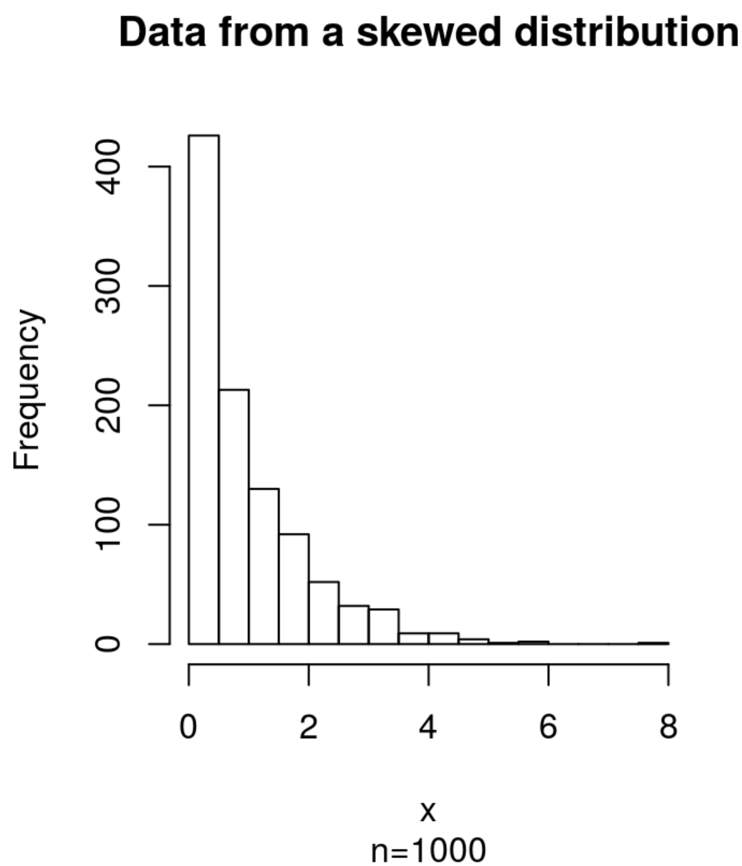
Runner data: What went wrong? (n=5)

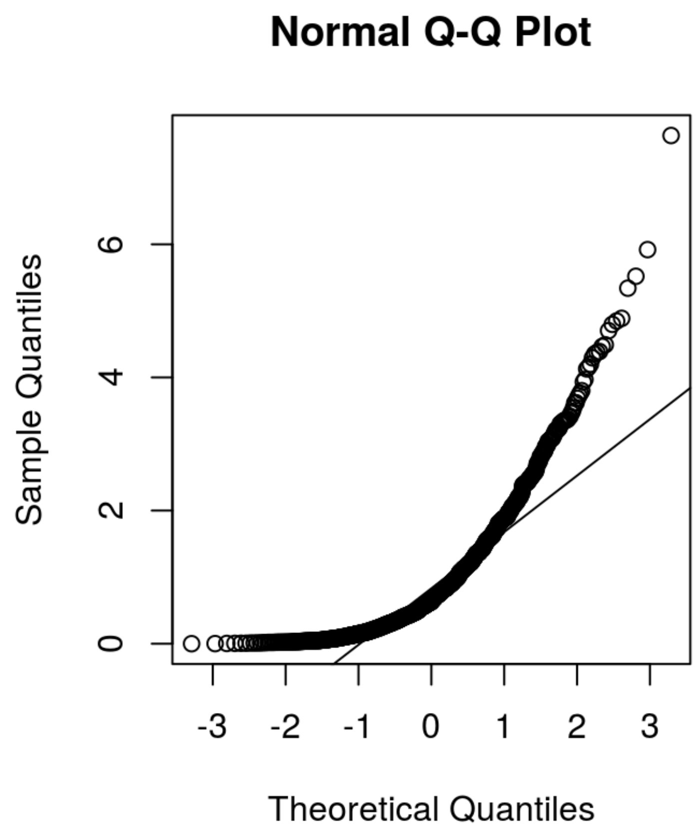


How to check Normality?

- Draw a histogram, check for symmetry and bell shape
- It is not clear to gauge if tails are heavier than Normal
- **Another diagnostic plot: Quantile-Quantile (Q-Q) plot**, also known as **Normal probability plot**
- Plot data against theoretical normal quantiles, see if they fall on a straight line

Q-Q Plot when normality does not hold





Q-Q plot for normally distributed data