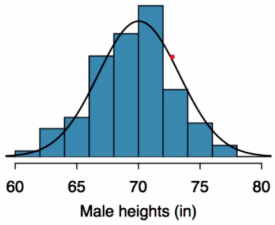
Plotting normal probability

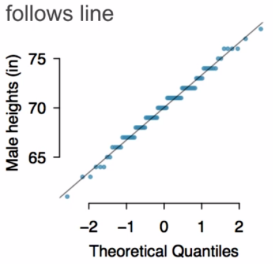
Histogram with best fitting normal curve

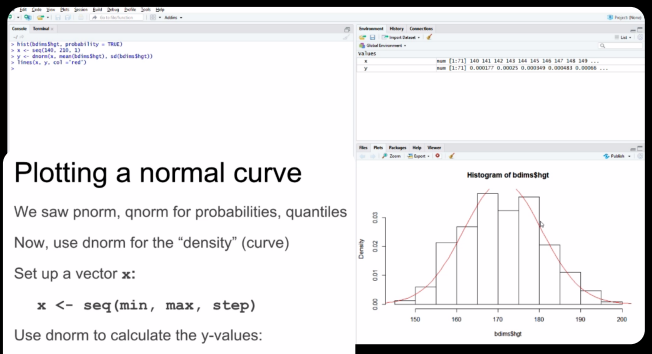
Distribution is normal if curve follows data



"Normal probability plot" or "quantile-quantile plot"

Distribution is normal if data follows line





Plotting a normal curve

We saw pnorm, qnorm for probabilities, quantiles

Now, use dnorm for the "density" (curve)

Set up a vector `x`:

```
x <- seq(min, max, step)
```

Use dnorm to calculate the y-values:

```
curve <- dnorm(x, mu, sigma)
```

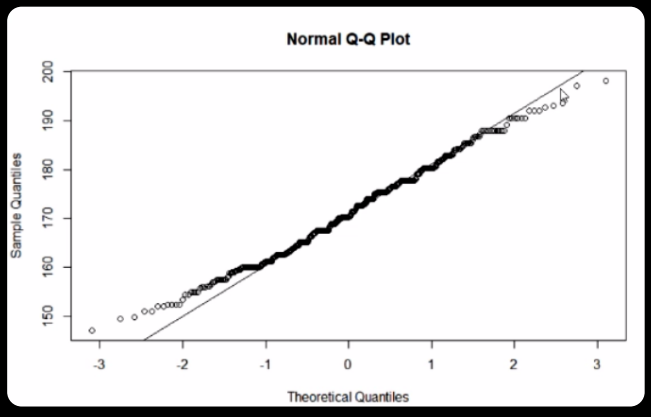
Use lines to add it to the plot:

```
lines(x, curve)
```

Creating a quantile-quantile plot

1. Sort N observations
2. Generate N values from a normal distribution:
   a. Calculate N equally-spaced probabilities in 0-1
   b. Find value corresponding to each probability
3. Sort normal distribution values
4. Plot sorted observations against sorted normal distribution values

```
qqnorm(bdims$hgt)
qqline(bdims$hgt)
```



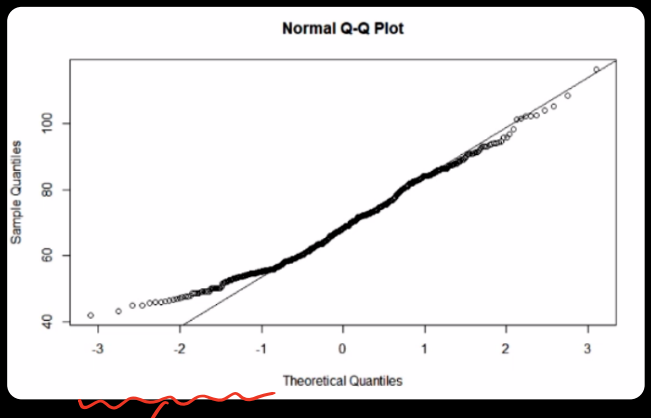
Quantile-quantile plot in R

In R, we have built in functions to handle the q-q plot:

- `qqnorm(data)` - plot the q-q points
- `qqline(data)` - plot the theoretical line the q-q points should follow *if* the distribution is normal

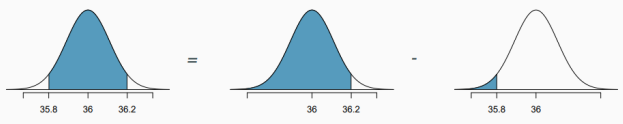
Note that some deviation from the line is typical, especially near the tails of the distribution

```
> qqline(bdims$wgt)
> hist(bdims$wgt, probability = TRUE)
```



What percent of bottles pass the quality control inspection?

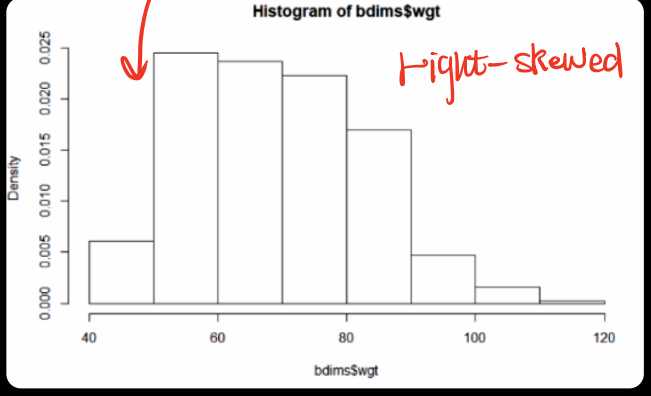
(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) *93.12%*
(e) 96.56%



$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

$$P(35.8 < X < 36.2) = P(-1.82 < Z < 1.82) = 0.9656 - 0.0344 = 0.9312$$



Parameter estimation

Want to know "parameters" of the true population

- Median, mean, standard deviation, etc.

But it's impractical to study whole population

- Use sample statistics as estimates

Sample statistics vary from sample to sample

- Distribution of estimates: *sampling distribution*
- Quantifying its variability: margin of error

Example

You are interested in studying blood calcium levels in elderly patients. Assume that in this population, the mean blood calcium level is 3.40 mmol/L and the standard deviation is 0.55 mmol/L. If you take a sample of 25 measurements and compute the mean:

- What are the mean and standard deviation of the sample mean?
- What is the probability that an individual observation is less than 3.30 mmol/L?
- What is the probability that your sample mean is less than 3.30 mmol/L?

- What are the mean and standard deviation of the sample mean?

Mean: same as population mean, 3.40 mmol/L

SD: population SD / sqrt(sample size), 0.11 mmol/L

$\frac{0.55}{\sqrt{25}}$

- What is the probability that an individual observation is less than 3.30 mmol/L?

```
> pnorm(3.30, 3.40, 0.55)
[1] 0.4278627
```

- What is the probability that the sample mean is less than 3.30 mmol/L?

```
> pnorm(3.30, 3.40, 0.11)
[1] 0.1816511
```

- What if instead you take a sample of 100 measurements?

The standard error would be 0.055 instead of 0.11.

95% Confidence Interval

Roughly: point estimate ± 2 × standard error

Example:

A random sample of 50 college students were asked: how many exclusive relationships have you been in? Sample mean was 3.2, sample standard deviation was 1.74.

Estimate the true population mean.

$$SE = \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}} \approx \frac{1.74}{\sqrt{50}} \approx 0.25$$

$$3.2 \pm 2 \times 0.25 = (3.2 - 0.5, 3.2 + 0.5) = (2.7, 3.7)$$

Sampling distributions

The *sampling distribution* of a quantity calculated from a sample (such as a sample mean) is the distribution of that quantity over all possible samples.

Central Limit Theorem

Given:

- D, a distribution of any shape
- D has mean μ and standard deviation σ

Then:

- Take samples of size M from D, find means
- The distribution of these means is well approximated by: $N\left(\mu, \frac{\sigma}{\sqrt{M}}\right)$

The standard deviation of the sampling distribution is called the standard error.

Central Limit Theorem

Only applies when:

1. Observations are independent
   - E.g., simple random sampling with sample size < 10% of population
2. Either the distribution sampled from is normal, or if it's skewed, the sample size is large
   - Rule of thumb: for moderately skewed distributions, sample size > 30
   - For extremely skewed distributions, consider a transformation

Confidential Interval

Confidence intervals

Context: estimating a population mean

Goal: produce an estimate that incorporates both our *best guess* for the population mean and a *margin of error* that quantifies how *specific* and how *uncertain* our guess is

This estimate takes the form of an interval: a lower and upper bound that we believe encloses the true population mean

95% confident in this interval

95% Confidence Interval

Why point estimate ± 2 × standard error?

Central Limit Theorem ⇒
Sample means are normally distributed

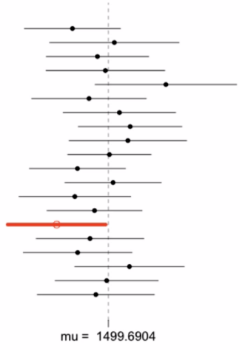
Normal distribution ⇒
≈95% of values within ± 2 standard errors

Slightly more precise:
point estimate ± 1.96 × standard error

What does 95% confident mean?

95% confidence is a statement about all possible samples of the same size and type.

If we took many samples (instead of just the 1 we took), about 95% of the confidence intervals would contain the true mean

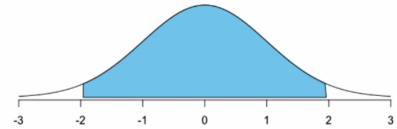


Changing the confidence level

We can find a multiplier for any confidence level:

point estimate ± z* × standard error

z* is called a *critical value*. 95% confidence:




```
> qnorm(0.025)  > qnorm(0.975)
[1] -1.959964   [1] 1.959964
```

95% ⇒ 1 − (100% − 95%)/2
⇒ 97.5%

What does 95% confidence *not* mean?

Incorrect	Correct
• 95% of observations are within the interval • 95% of possible sample means are within the interval	• Confidence intervals say nothing about individuals • 95% confidence is a statement about the relationship between *this specific sample* and population mean

only population

Reduce standard error
Improve sample size

Tradeoff: specificity vs. certainty

A narrower interval means we are making a more *specific* estimate.

But a narrower interval comes with a lower degree of *certainty* (confidence level).

The only way to win on both fronts is to reduce the standard error:

- More data (larger sample)
- Better measurement (lower SD)

Sample size for a margin of error

Studies suggest that the standard deviation of IQ scores of 3-year-olds is 18 points. How many children should we sample for a 98% confidence interval with a margin of error ≤ 4 points?

$$\text{margin} = z^* \times \text{SE} = z^* \times \frac{\sigma}{\sqrt{n}}$$

$$n = \left(z^* \times \frac{\sigma}{\text{margin}}\right)^2$$

```
> (qnorm(0.99) * 18 / 4)^2
[1] 109.5909
```

Critical value

margin = critical value × standard deviation / √sample size

(qnorm((1 − confidence)/2))

$$m = z^* \times \frac{\sigma}{\sqrt{n}}$$

$$\frac{m\sqrt{n}}{z^*} = \sigma$$