Unit 2: Foundations for Inference

3. The Normal Distribution and more on the Central Limit Theorem

(2.6)

2/10/2020

Recap from last time

1. Larger samples give us more precision

2. The Central Limit Theorem says that the Null distribution will generally approach the Normal distribution

3. Using theoretical distributions (instead of shuffled random distributions) makes statistical measures lossless compression

Key ideas

1. We are really thinking about three distributions: the sample, the population, and the test statistic

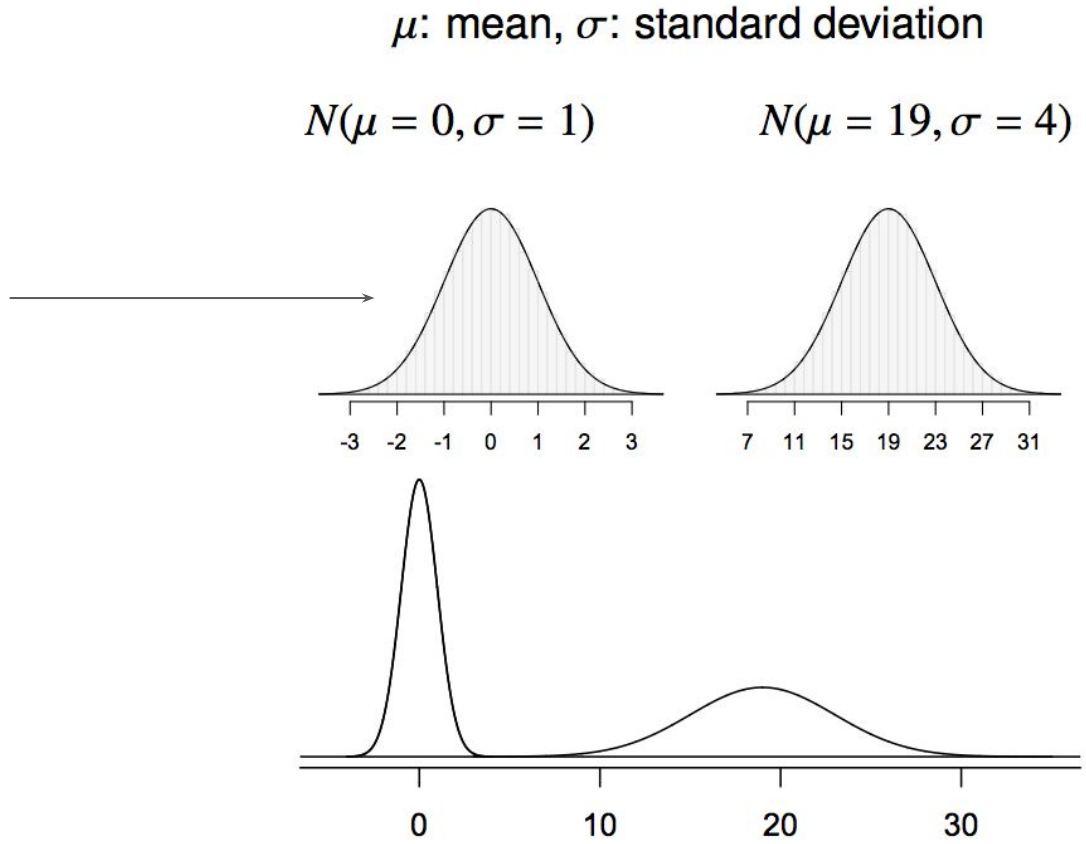
2. We can use Z-scores to compare points on two different normal distributions

3. We can use Quantile-Quantile Plots to check for Normality

A reminder about the central limit theorem

Different Normal Distributions

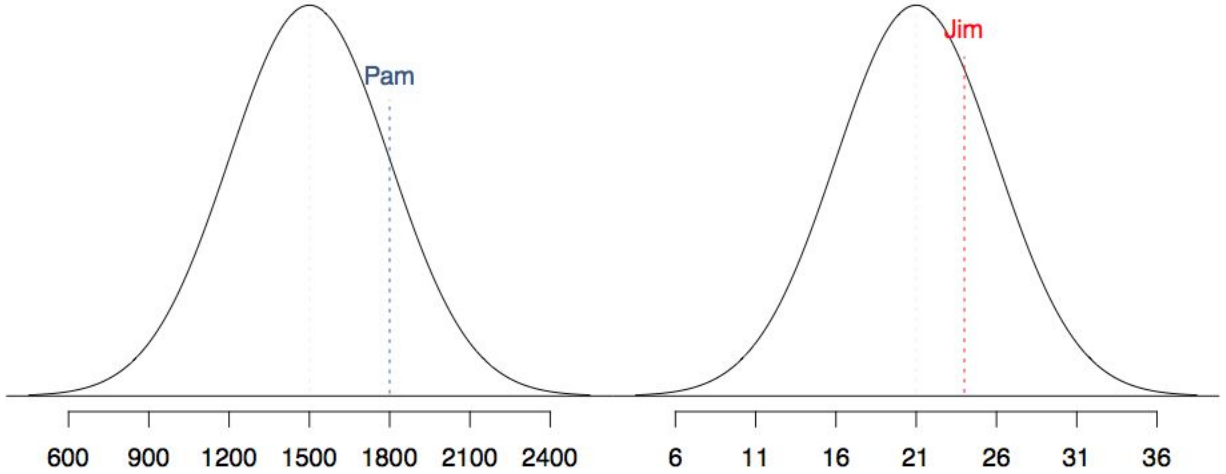
Standard Normal Distribution



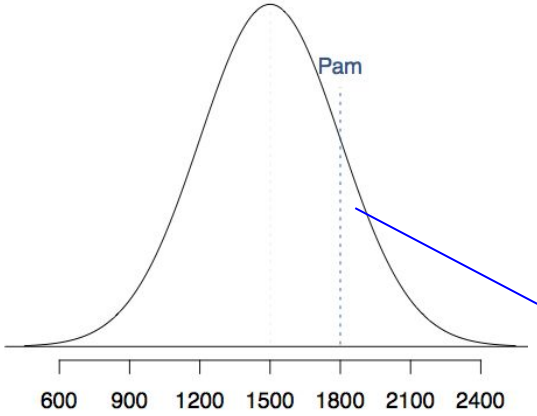
Comparing samples from two normal distributions

A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?

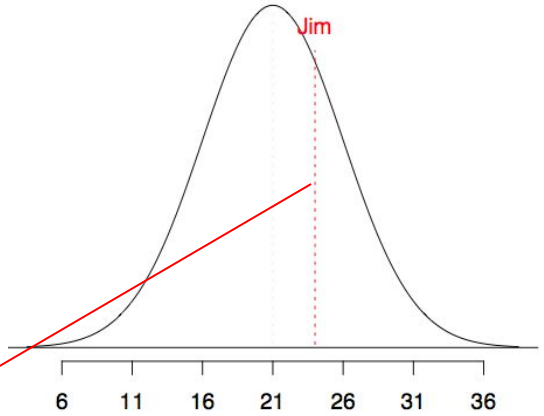
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5.



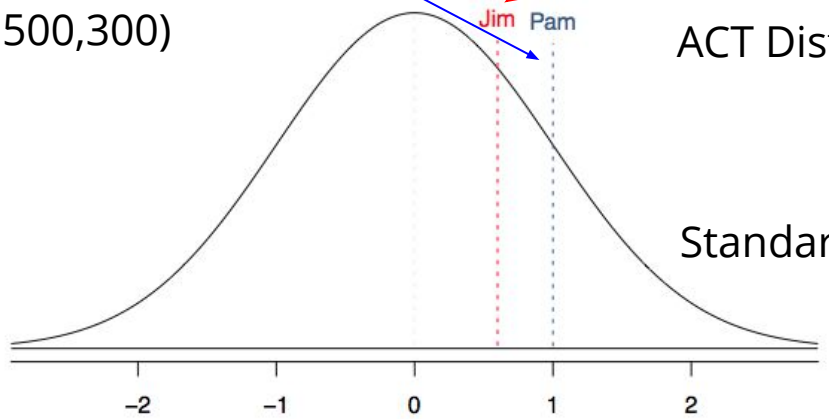
We can map different Normal Distributions onto the Standard Normal



SAT Distribution: N(1500,300)



ACT Distribution: N(21,5)



Standard Normal: N(0,1)

Z-score: Number of Standard Deviations above the mean

These are called standardized scores, or Z-scores.

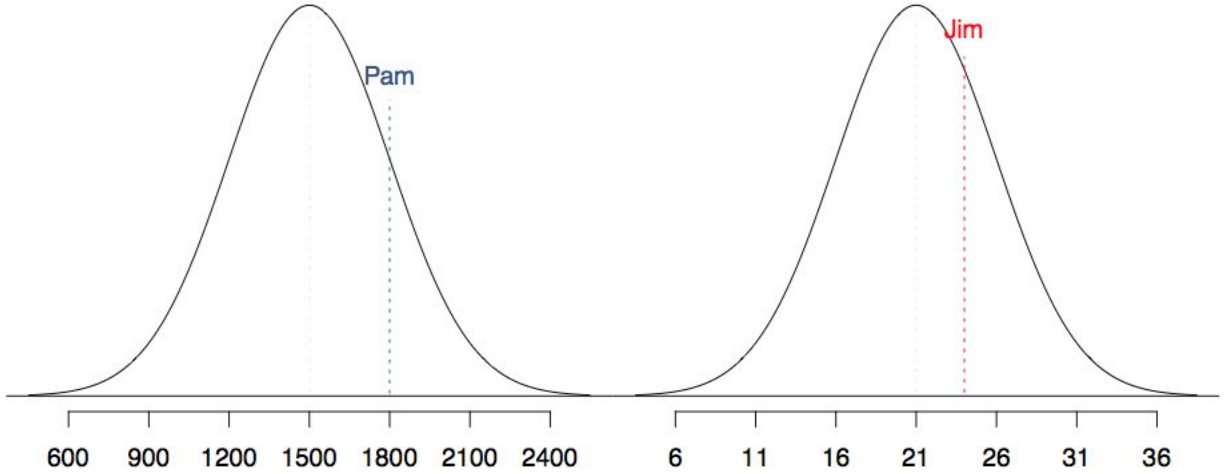
- Z-score of an observation is the number of standard deviations it falls above or below the mean.

  Z = (observation - mean) / SD

Comparing samples from two normal distributions

We can't just compare these two raw scores. But, we *can* compare how many standard deviations beyond the mean each observation is.

- Pam's score is (1800 - 1500) / 300 = 1 standard deviation above the mean.
- Jim's score is (24 - 21) / 5 = 0.6 standard deviations above the mean.



Z-score: Number of Standard Deviations above the mean

These are called standardized scores, or Z-scores.

- Z-score of an observation is the number of standard deviations it falls above or below the mean.

  Z = (observation - mean) / SD

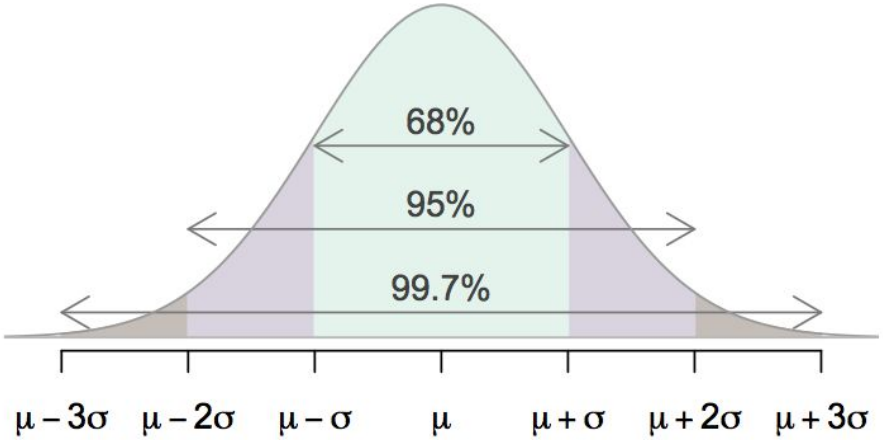
- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

The 68-95-99.7 Rule

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Practice Question 1: Quality Control

At the Heinz ketchup factory, the amount of ketchup that goes into the bottle is supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz.

Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection.

What percent of bottles have less than 35.8 ounces of ketchup?

$Z(35.8) = (35.8 - 36)/.11 \sim -1.82$

Since ~95% of the distribution falls within 2SD on either side of the mean, we should expect it to be a little bit more than 2.5%

Practice Question 2: Quality Control

What percent of bottles pass the quality control inspection?

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

Practice Question 2: Quality Control

What percent of bottles pass the quality control inspection?

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

Aside: Don’t worry about probability tables. We live in 2020

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5

use

pnorm(Z)

`pnorm` and `qnorm`

`pnorm(q, mean, sd)`: get the probability of the normal distribution with mean and standard deviation (sd) associated with a given quantile

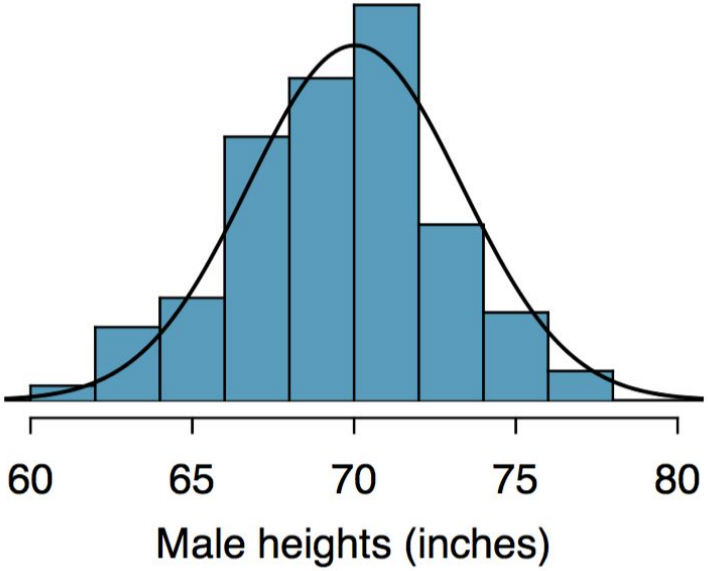
```
pnorm(1.96, 0, 1) = .975
```

`qnorm(p, mean, sd)`: get the quantile of the normal distribution with mean and standard deviation (sd) associated with a given probability

```
qnorm(.975, 0, 1) = 1.96
```

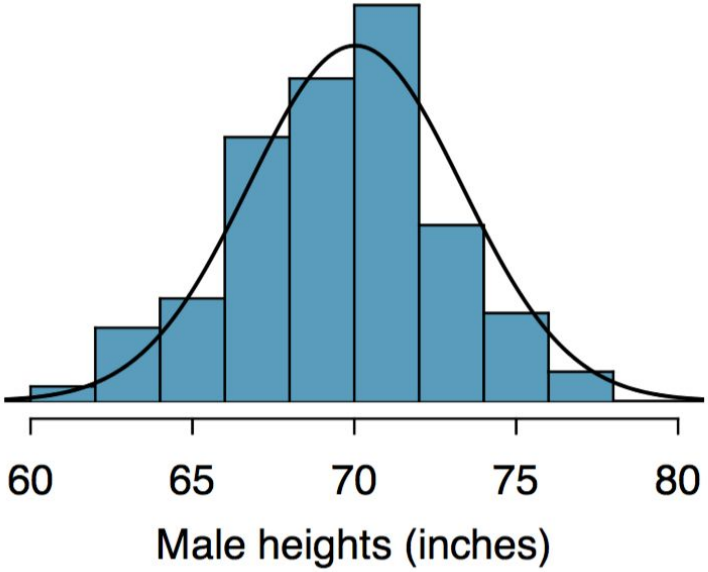
How do you know if a distribution is Normal?

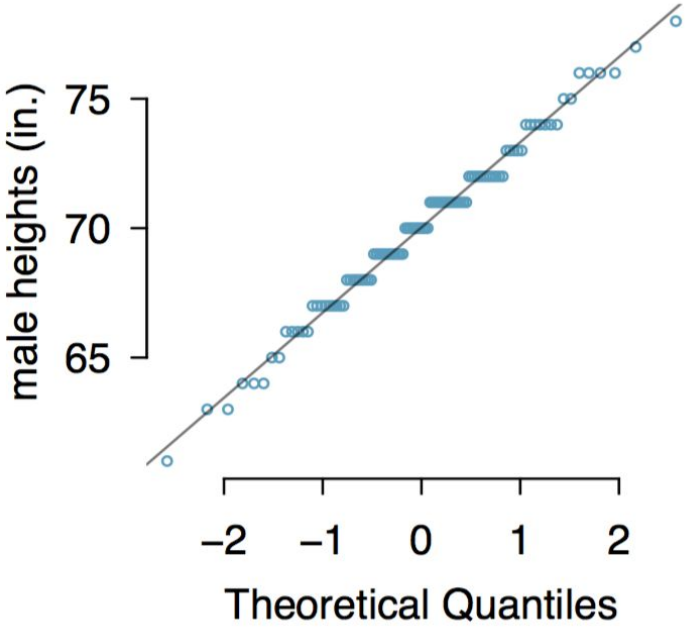
Draw a normal distribution over it, see how good it looks.



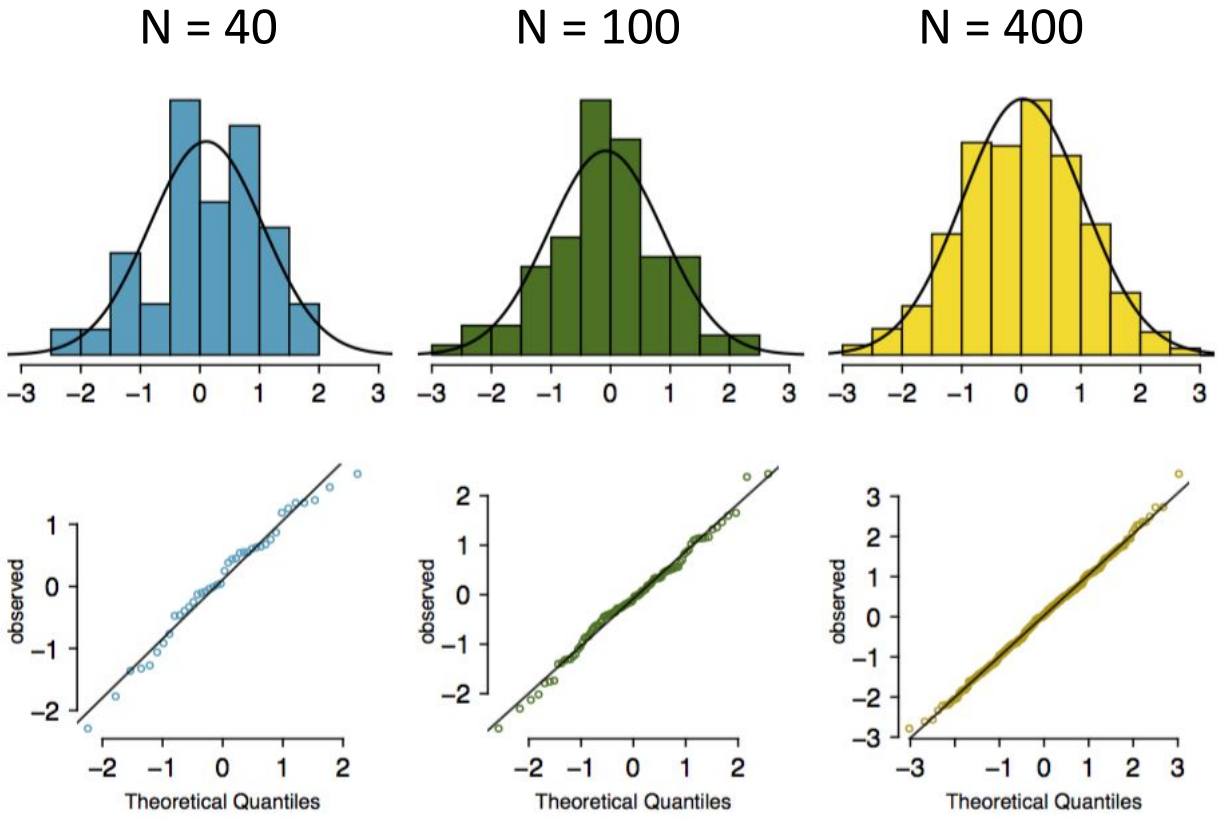
How do you know if a distribution is Normal?

An easier plot to look at is a **Quantile-Quantile (QQ) Plot**

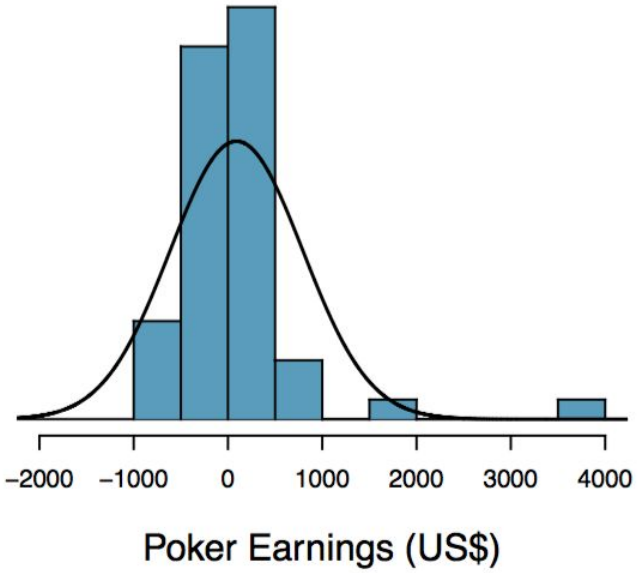


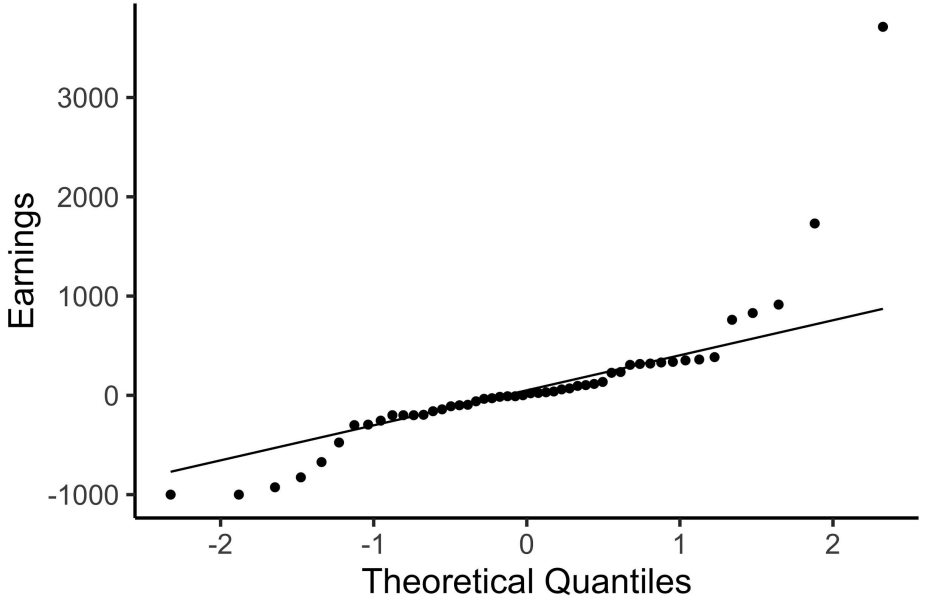


Poorly sampled Normal plots show non-systematic deviations



Non-normal plots show systematic deviations






```
  ggplot(poker, aes(sample = winnings))
+
     geom_qq() +
     geom_qq_line()
```

Practice Question 3: Which of the following is **false**

1. Majority of Z-scores in a right skewed distribution are negative.
2. In skewed distributions the Z-score of the mean might be different than 0.
3. For a normal distribution, IQR is less than 2 x SD.
4. Z-scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Practice Question 3: Which of the following is **false**

1. Majority of Z-scores in a right skewed distribution are negative.
2. **In skewed distributions the Z-score of the mean might be different than 0.**
3. For a normal distribution, IQR is less than 2 x SD.
4. Z-scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Key ideas

1. We are really thinking about three distributions: the sample, the population, and the test statistic

2. We can use Z-scores to compare points on two different normal distributions

3. We can use Quantile-Quantile Plots to check for Normality